ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO TỔNG KẾT

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

NĂM 2024

PHƯƠNG THỨC TIẾP CẬN LINH HOẠT NHẰM DỰ ĐOÁN RỜI BỎ BẰNG PHÂN LỚP DỮ LIỆU CHUỖI THỜI GIAN

Lĩnh vực khoa học: **Nghiên cứu ứng dụng**

Chuyên ngành: **Công nghệ thông tin**

Nhóm nghiên cứu:

TT	Họ tên	MSSV	Đơn vị	Nhiệm vụ	Điện thoại	Email
1	Nguyễn Duy Duẩn	K214061736	Khoa Hệ thống thông tin	Tham gia	0847468525	duannd21406@st.uel.edu.vn
2	Nguyễn Thúy Vy	K214061753	Khoa Hệ thống thông tin	Tham gia	0915753040	vynt21406@st.uel.edu.vn
3	Trần Sĩ Đan	K214061258	Khoa Hệ thống thông tin	Nhóm trưởng	0945439636	dants21406@st.uel.edu.vn

ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO TỔNG KẾT

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN

NĂM 2024

PHƯƠNG THỨC TIẾP CẬN LINH HOẠT NHẰM DỰ ĐOÁN RỜI BỎ BẰNG PHÂN LỚP DỮ LIỆU CHUỖI THỜI GIAN

Đại diện nhóm nghiên cứu	**Giảng viên hướng dẫn**	**Chủ tịch Hội đồng**
(Ký, họ tên)	*(Ký, họ tên)*	*(Ký, họ tên)*
		
	Lê Thị Kim Hiền	

Lãnh đạo Khoa/Bộ môn/Trung tâm

(Ký, họ tên)

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

DANH MỤC NHỮNG TỪ VIẾT TẮT

Từ viết tắt	Chữ viết đầy đủ
ADASYN	Adaptive Synthetic Sampling
AUC	Area Under the Curve
CLTV	Customer Lifetime Value
CNN	Convolutional Neural Network
COTE	Collective of Transformation-based Ensemble
DNN	Deep Neural Network
DT	Decision Tree
EMG	Electromyography, phương pháp đo điện hoạt động sản sinh bởi cơ bắp
ESN	Echo State Network
FCN	Fully-Connected Network
LR	Linear Regression
LSTM	Long Short Term Memory
MLP	Multi Layer Perceptron
ResNet	Residual Network
SLP	Single Layer Perceptron
SVM	Support Vector Machine
TSC	Time Series Classification
TSK means	Time Series K-means
TSOC	Time Series with Ordered Categories Customer Lifetime Value

TÓM TẮT

Vấn đề về giữ chân khách hàng đã trở thành một trong những ưu tiên hàng đầu của các doanh nghiệp do khách hàng rời bỏ luôn đi kèm với thất thoát về doanh thu. Vì vậy, việc thấu hiểu hành vi của khách hàng và kịp thời can thiệp quyết định rời bỏ là vô cùng quan trọng. Nghiên cứu giải pháp giữ chân khách hàng không chỉ gói gọn trong dự đoán khách hàng rời bỏ mà còn là xác định các đặc điểm của khách hàng và nguyên nhân khách hàng rời bỏ. Trong bài nghiên cứu này, nhóm tác giả đề xuất phương pháp tiếp cận mới mẻ đối với dự đoán rời bỏ sử dụng mô hình phân lớp chuỗi thời gian. Phương pháp tiếp cận linh hoạt này giúp sớm nhận diện quyết định rời bỏ của khách hàng, tìm ra các yếu tố ảnh hưởng đến quyết định rời bỏ và xác định thời điểm khách hàng quyết định rời bỏ. Nhóm nghiên cứu đề xuất mô hình phân lớp chuỗi thời gian giải thích được MiniRocket-SHAP, giúp dự đoán rời bỏ chính xác và giải thích kết quả dự đoán. Mô hình MiniRocket-SHAP được so sánh với mô hình phân lớp chuỗi thời gian LSTM kết hợp SLP và mô hình truyền thống RF-Static. Ngoài ra, các không gian đặc trưng khác nhau cũng được so sánh để đánh giá hiệu quả của phương pháp chọn đặc trưng từ giá trị Shapley của mô hình MiniRocket-SHAP. Kết quả nghiên cứu cho thấy mô hình MiniRocket-SHAP sử dụng các đặc trưng được chọn dựa trên giá trị Shapley đạt kết quả tốt nhất với chỉ số Accuracy là 0.95 và chỉ số F1 là 0.79.

Keyword: phân lớp chuỗi thời gian, học máy, học sâu, nền tảng âm nhạc trực tuyến, dự đoán rời bỏ, khách hàng rời bỏ, giải thích mô hình.

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1. Lý do lựa chọn đề tài

Trong nền kinh tế hiện nay, các doanh nghiệp trong thị trường cạnh tranh chủ yếu dựa vào lợi nhuận thu được từ khách hàng. Việc khách hàng rời bỏ có thể dẫn đến thất thoát lợi nhuận của doanh nghiệp, do đó dự đoán tỷ lệ khách hàng rời bỏ ngày càng nhận được sự quan tâm trong các tài liệu tiếp thị và quản lý trong thời gian qua. Tỷ lệ rời bỏ ảnh hưởng đến lợi nhuận của một công ty vì 65% doanh thu của công ty đến từ các khách hàng hiện tại, tạo ra một nhu cầu là phải xác định và ngăn chặn tình trạng khách hàng rời bỏ (S. Kumar Hegde và cộng sự, 2023). Trước đây, hiệu quả trong việc thu hút khách hàng được đánh giá bằng việc so sánh số lượng khách hàng rời bỏ với số lượng mới. Tuy nhiên, khi thị trường bão hòa do sự toàn cầu hóa của dịch vụ và cạnh tranh gay gắt, chi phí thu hút khách hàng mới đã tăng nhanh chóng, cao hơn chi phí giữ chân khác hàng (T. Verbraken và cộng sự, 2014). Tỷ lệ rời bỏ có tác động đáng kể đến giá trị trọn đời của khách hàng vì nó ảnh hưởng đến doanh thu trong tương lai của công ty cũng như thời gian cung cấp dịch vụ (Kavitha và cộng sự, 2020). Trong nhiều trường hợp hơn giả định, hoạt động tiếp thị có thể thành công xây dựng lượng người mua trung thành thay vì liên tục thu hút những người mới. Hoạt động này có thể đảm bảo thị phần có thể phòng thủ được và ngày càng tăng.

Nhằm mục đích quản lý khách hàng rời bỏ một cách hiệu quả, các công ty cần xây dựng một mô hình dự đoán hiệu quả và chính xác. Trong nhiều tài liệu nghiên cứu, kỹ thuật thống kê và khai thác dữ liệu đã được sử dụng để tạo lập các mô hình dự đoán. Một số nghiên cứu sử dụng học máy cho việc dự đoán, cụ thể các mô hình dự đoán được sử dụng như hồi quy logistic, SVM, Decision Tree như (Lalwani và cộng sự, 2022). Có thể nhận thấy rằng mặc dù các bài nghiên cứu này có đóng góp tích cực trong việc cải tiến các mô hình dự đoán tuy nhiên vẫn tồn tại nhiều hạn chế, nhất là đối với bài toán triển khai dự đoán rời bỏ thường xuyên nhằm hỗ trợ các chiến dịch giữ chân khách hàng. Các nghiên cứu trên sử dụng dữ liệu nghiên cứu ở dạng tĩnh, như tổng, trung bình,... của các dữ liệu trong quá khứ. Việc xây dựng mô hình dự đoán với dữ liệu tĩnh bỏ qua bản chất thay đổi theo thời gian của hành vi khách hàng đã làm giảm hiệu quả dự đoán và không tối ưu cho việc triển khai thường xuyên, cập nhật các biến động

mới nhất trong hành vi khách hàng. Vì đó, dữ liệu chuỗi thời gian sẽ phù hợp hơn với tính chất hành vi khách hàng. Tuy nhiên, các nghiên cứu trước đây vẫn tồn tại ba hạn chế lớn khác. Thứ nhất, việc tác giả thiết lập thực nghiệm chưa phù hợp với bản chất của dữ liệu chuỗi thời gian và chưa phản ánh được khả năng triển khai trong thực tế do tính chất thời gian của dữ liệu, chuỗi thời gian từ hai khung thời gian khác nhau trong thực tế có thể có tính chất khác nhau. Ví dụ, nếu chuỗi thời gian thể hiện doanh thu của một doanh nghiệp có xu hướng tăng ở đầu tháng và xu hướng giảm ở cuối tháng, việc huấn luyện và kiểm thử mô hình dự đoán dựa vào dữ liệu ở đầu tháng sẽ bỏ qua các xu hướng ở cuối tháng, vì vậy làm giảm khả năng tổng quát hóa của mô hình dự đoán. Thứ hai, phương pháp dự đoán rời bỏ truyền thống thường được sử dụng để dự đoán theo chu kỳ tháng, tức là dự đoán được thực hiện khi kết thúc một tháng và kết quả được sử dụng cho chiến dịch giữ chân khách hàng ở tháng tiếp theo. Điều này dẫn đến việc doanh nghiệp không thể can thiệp kịp thời và đưa ra chiến lược hiệu quả để giữ chân những khách hàng đã quyết định rời đi vào đầu tháng trước đó. Theo Alboukaey và cộng sự (2020), việc sử dụng một mô hình dự đoán rời bỏ thường xuyên ở mức độ ngày hoặc tuần giúp thiết kế chiến dịch giữ chân vào bất kỳ thời điểm nào trong tháng tùy thuộc vào xu hướng mới cập nhật chứ không phải xu hướng cũ từ tháng trước. Thứ ba, đối với các nghiên cứu sử dụng phân lớp chuỗi thời gian, phần lớn các mô hình hiện đại là không giải thích được (hay còn gọi là black-box model) và chưa có nhiều bài nghiên cứu đề xuất hướng phân tích mở rộng từ kết quả phân loại nhằm giúp doanh nghiệp đưa ra chiến lược giữ chân khách hàng. Theo Theissler và cộng sự (2022), khi các mô hình trở nên chính xác và phức tạp hơn, việc thiếu khả năng giải thích hoặc diễn giải mô hình là một trong những thách thức chính của nghiên cứu máy học. Thách thức như vậy có thể ngăn cản việc sử dụng ML trong các ứng dụng yêu cầu các quyết định có thể hiểu được. Những thách thức như vậy đặt ra yêu cầu là phải có một mô hình không chỉ đảm bảo tính chính xác mà còn linh hoạt và nhanh chóng đưa ra kết quả, giúp doanh nghiệp kịp thời có kế hoạch cho việc giữ chân khách hàng.

Nghiên cứu "***Phương pháp tiếp cận linh hoạt dự đoán khách hàng rời bỏ sử dụng phân lớp chuỗi thời gian***" đề xuất hướng tiếp cận linh hoạt nhằm dự đoán khách hàng rời bỏ sử dụng phân lớp chuỗi thời gian. Nghiên cứu hướng đến khắc phục những nhược điểm của các nghiên cứu trước đó, nhấn mạnh sự thay đổi hành vi của khách

hàng theo tuần, chú trọng vào tính kịp thời trong việc phân tích dự đoán và đưa ra kế hoạch giữ chân khách hàng hiệu quả. Đồng thời, nhóm nghiên cứu đề xuất một mô hình mới, gọi là MiniRocket-SHAP giúp dự đoán rời bỏ sử dụng phân lớp chuỗi thời gian và xác định nguyên nhân, thời điểm khách hàng quyết định rời bỏ.

1.2. Mục tiêu

Đề tài hướng đến các mục tiêu chung sau:

Về mặt kỹ thuật:

- Ứng dụng phân lớp chuỗi thời gian vào dự đoán hành vi rời bỏ của khách hàng.
- Đề xuất cách tiếp cận sử dụng phân lớp chuỗi thời gian để dự đoán rời bỏ chính xác, giải thích được và ứng dụng được để đưa ra các chiến lược.

Về mặt thực tiễn:

Đề xuất một phương pháp linh hoạt cho doanh nghiệp nhằm dự đoán chính xác khách hàng rời bỏ để đưa ra những chiến lược giữ chân phù hợp và kịp thời.

Cụ thể, đề tài sẽ bao gồm những nội dung chính sau:

Nội dung 1: Nghiên cứu, tìm kiếm dữ liệu và phân tích về hành vi khách hàng về giao dịch đăng ký, gia hạn; các hành vi sử dụng dịch vụ

- Dựa trên mục tiêu nghiên cứu, nhóm tiến hành tìm kiếm bộ dữ liệu, khảo sát các đặc trưng phù hợp về hành vi có thể được sử dụng cho phân loại chuỗi thời gian.
- Tiến hành thu thập và tiền xử lý bộ dữ liệu nhằm phù hợp để thực nghiệm.

Nội dung 2: Nghiên cứu và xây dựng mô hình phù hợp cho dự đoán rời đi và thực nghiệm

- Thực hiện khảo lược những nghiên cứu trước, tìm hiểu về các mô hình được sử dụng trong dự đoán rời bỏ.
- Tiến hành xây dựng mô hình theo những yêu cầu đặt ra.
- Từ tập dữ liệu đã được thu thập và xử lý ở Nội dung 1, nhóm tiến hành thử nghiệm mô hình đã xây dựng.

Nội dung 3: Đánh giá mức độ hiệu quả

- Tiến hành so sánh kết quả thực nghiệm giữa các mô hình.
- Nghiên cứu và xây dựng khung chuẩn để đánh giá mức độ hiệu quả của hệ thống và tìm ra biện pháp tối ưu.

1.3. Đối tượng và phạm vi nghiên cứu

Đối tượng: Các dữ liệu mô hình doanh thu đăng ký trên nền tảng âm nhạc, bao gồm dữ liệu về thanh toán giao dịch đăng ký và gia hạn có ghi nhận thời gian phát sinh; các nhật ký người dùng được ghi nhận trên nền tảng âm nhạc trực tuyến.

Phạm vi:

- *Phạm vi không gian:* Thực hiện với bộ dữ liệu từ KKBOX, một nền tảng âm nhạc trực tuyến, ghi lại hành vi của khách hàng khi sử dụng dịch vụ phát nhạc này.
- *Phạm vi thời gian:* 6 tháng.

1.4. Phương pháp nghiên cứu

Phương pháp nghiên cứu định tính: Khảo sát những đặc trưng về hành vi khách hàng trong sử dụng nền tảng âm nhạc trực tuyến (theo mô hình doanh thu đăng ký) để nắm bắt được sơ lược nhằm phục vụ cho quá trình trích xuất và biến đổi đặc trưng của dữ liệu và thiết kế mô hình phù hợp.

Phương pháp thực nghiệm: Tiến hành thực nghiệm trên mô hình chính và một số mô hình khác để so sánh với nhau nhằm đánh giá mức độ hiệu quả, đưa ra phương hướng cải thiện và phát triển.

Phương pháp nghiên cứu định lượng: Thống kê dữ liệu và số liệu nhằm đưa ra kết quả trực quan.

1.5. Kết cấu bài nghiên cứu

Với đề tài nghiên cứu, báo cáo tổng kết được tổ chức như sau:

Chương 1: Giới thiệu tổng quan đề tài

Nhóm tiến hành khái quát hóa bài nghiên cứu như: Lý do chọn đề đề tài, tổng quan tình hình nghiên cứu, mục tiêu nghiên cứu, đối tượng và phạm vi; phương pháp

nghiên cứu; bố cục và ý nghĩa thực tiễn của đề tài

Chương 2: Cơ sở lý thuyết, kỹ thuật liên quan đến nghiên cứu

Báo cáo đưa ra các lý thuyết cơ bản, phương pháp được được sử dụng trong bài nghiên cứu. Đồng thời, khái quát những nghiên cứu trước đó.

Chương 3: Mô hình nghiên cứu đề xuất

Ở chương này, các mô hình mà nhóm kế thừa được thảo luận và đánh giá cùng một số bài nghiên cứu có liên quan trực tiếp. Từ những đánh giá trên, sau khi rút ra được những điểm yếu của các mô hình trước đó, nhóm tiến hành xây dựng một mô hình đề xuất mới được chia thành các giai đoạn và được giải thích cụ thể.

Chương 4: Kết quả thực nghiệm và đánh giá kết quả.

Khi hoàn thành mô hình ở chương 3, nhóm tiến hành thực nghiệm toàn bộ theo mô hình đã đề xuất từ bước mô tả dữ liệu, quá trình tiền xử lý, chọn lọc thuộc tính, xây dựng thuộc tính, đưa vào xử lý trong mô hình chính. Sau khi có kết quả, nhóm tiến hành phân tích, đánh giá kết quả.

Chương 5: Thảo luận và đề xuất

Chương này đưa ra những thảo luận về kết quả của toàn bộ nghiên cứu, những khía cạnh của mô hình đã đề ra, suy xét tính hiệu quả của nó, từ đó đề xuất sử dụng mô hình cho các doanh nghiệp.

Chương 6: Kết luận và hướng phát triển

Nhóm rút ra kết luận về những đóng góp của bài nghiên cứu trên khía cạnh kỹ thuật lẫn kinh doanh, nhìn nhận một số hạn chế còn tồn tại và đề xuất hướng phát triển mở rộng trong tương lai.

Tóm tắt chương 1

Trong chương 1, nhóm đã trình bày tổng quan về nghiên cứu về lý do lựa chọn, mục tiêu mà nhóm hướng đến khi thực hiện nghiên cứu; đối tượng và phạm vi nghiên cứu; các phương pháp sử dụng và cuối cùng nêu kết cấu toàn bài nghiên cứu.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Các khái niệm liên quan

2.1.1. Khách hàng rời bỏ

Định nghĩa được phát hiện bởi Berson và cộng sự (2000) cho rằng "khách hàng rời bỏ" là quá trình những người đăng ký chuyển đổi từ một nhà cung cấp dịch vụ. Việc rời bỏ có thể chủ động - cố ý, luân phiên - ngẫu nhiên hoặc thụ động - không tự nguyện. Việc khách hàng rời bỏ là thước đo phổ biến về số lượng khách hàng bị mất. Khách hàng rời bỏ, được định nghĩa là xu hướng khách hàng ngừng giao dịch với một công ty trong một khoảng thời gian nhất định, đã trở thành một vấn đề quan trọng và là một trong những thách thức chính mà nhiều công ty trên toàn thế giới đang phải đối mặt (Chandar và cộng sự, 2006). Hay một định nghĩa khác cho rằng rời bỏ là khi một khách hàng ngưng sử dụng dịch vụ hoặc hủy đăng ký của họ (Gold, 2020).

2.1.2. Dự đoán khách hàng rời bỏ

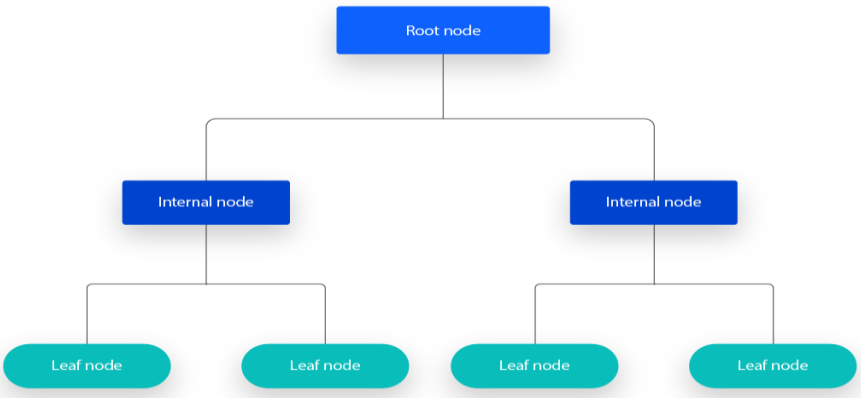
Để thành công, một dịch vụ phải nỗ lực để tối thiểu hóa vấn đề rời bỏ. Nếu vấn đề này không được giải quyết liên tục và chủ động thì sản phẩm dịch vụ sẽ không phát huy hết tiềm năng của nó (Gold, 2020). Một cách để quản lý việc rời bỏ khách hàng là dự đoán những khách hàng nào có nhiều khả năng rời bỏ nhất và sau đó nhắm mục tiêu khuyến khích những khách hàng đó để khiến họ ở lại (Pustokhina và cộng sự, 2021; Neslin và cộng sự, 2006 và Lalwani và cộng sự, 2022). Cách tiếp cận này cho phép công ty tập trung nỗ lực vào những khách hàng thực sự có nguy cơ rời bỏ và có khả năng tiết kiệm số tiền sẽ bị lãng phí khi cung cấp các ưu đãi cho những khách hàng không cần chúng. Tuy nhiên, cách tiếp cận này giả định rằng việc rời bỏ khách hàng có thể được dự đoán với độ chính xác chấp nhận được.

Nhận thấy tầm quan trọng của việc dự đoán khách hàng rời bỏ, nhiều công ty lớn thực hiện các mô hình dự đoán nhằm phát hiện những động thái rời đi của khách hàng. Những kỹ thuật về khai phá dữ liệu ngày càng được áp dụng rộng rãi (Burez & Van den Poel, 2009).

2.2. Cơ sở lý thuyết

2.2.1. Cây quyết định

Cây quyết định là một thuật toán học tập có giám sát không tham số, được sử dụng cho cả nhiệm vụ phân loại và hồi quy (Breiman và cộng sự, 1984). Nó có cấu trúc cây phân cấp, bao gồm một nút rễ, các nhánh, nút bên trong và nút lá (Hình 2-1).



Hình 2-1. Cây quyết định

Hai phương pháp chọn thuộc tính tốt nhất tại mỗi nút được sử dụng nhiều nhất là độ lợi của thông tin (Information Gain) và tạp chất Gini (Gini Impurity), đóng vai trò là tiêu chí phân tách phổ biến cho các mô hình cây quyết định. Chúng giúp đánh giá chất lượng của từng điều kiện thử nghiệm và mức độ có thể phân loại mẫu thành một lớp.

Phương pháp độ lợi thông tin: Để chọn tính năng tốt nhất để phân tách và tìm cây quyết định tối ưu, nên sử dụng thuộc tính có lượng entropy nhỏ nhất. Độ lợi thông tin thể hiện sự khác biệt về entropy trước và sau khi phân chia trên một thuộc tính nhất định. Thuộc tính có độ lợi thông tin cao nhất sẽ tạo ra sự phân chia tốt nhất vì nó thực hiện công việc phân loại dữ liệu huấn luyện tốt nhất theo phân loại mục tiêu của nó. Với entropy là một khái niệm bắt nguồn từ lý thuyết thông tin, đo lường tạp chất của các giá trị mẫu, được định nghĩa bằng công thức:

$$Entropy(S) = -\sum_{i}^{C} p(i) log_2 p(i)$$

Trong đó:

- S đại diện cho tập dữ liệu mà $entropy$ được tính toán.
- i đại diện cho các lớp cụ thể.
- $p(i)$ đại diện cho tỷ lệ các điểm dữ liệu của lớp i với tổng số điểm dữ liệu trong tập hợp S.

Độ lợi thông tin được tính bằng công thức:

$$Information\ Gain\ (S, A)\ =\ E(S)\ -\ \sum_{a\ \epsilon\ A} \frac{|S_a|}{|S|} E(S_a)$$

Trong đó:

- E đại diện cho giá trị $entropy$.
- a đại diện cho thuộc tính cụ thể.
- $\frac{|S_a|}{|S|}$ biểu thị tỷ lệ của các giá trị thuộc tính a cụ thể với số lượng giá trị trong tập dữ liệu.

Phương pháp tạp chất Gini: xác suất phân loại không chính xác điểm dữ liệu ngẫu nhiên trong tập dữ liệu nếu nó được dán nhãn dựa trên phân phối lớp của tập dữ liệu và được thể hiện dưới công thức:

$$Gini\ Impurity(S)\ =\ 1\ -\ \sum_{i}^{C} (p_i)^2$$

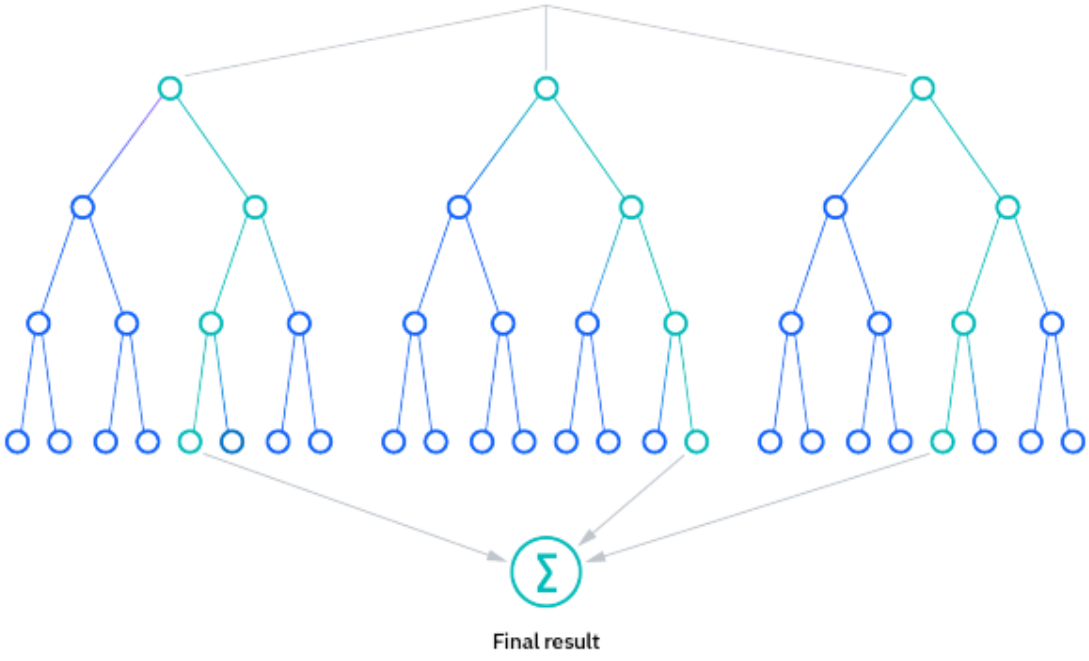
Trong đó:

- S đại diện cho tập dữ liệu mà $entropy$ được tính toán.
- i đại diện cho các lớp cụ thể.
- $p(i)$ đại diện cho tỷ lệ các điểm dữ liệu của lớp i với tổng số điểm dữ liệu trong tập hợp S.

2.2.2. Rừng ngẫu nhiên

Rừng ngẫu nhiên là một loại phương pháp học tập kết hợp nhiều cây quyết định để cải thiện độ chính xác và độ bền vững của mô hình dự đoán. Khái niệm về rừng ngẫu nhiên được đưa ra lần đầu tiên bởi Ho (1995), người đề xuất rằng mỗi cây nên được xây

dựng từ một tập con ngẫu nhiên của không gian đặc trưng để tạo ra các cây khác nhau (Breiman, 2001) sau đó giới thiệu một phiên bản phổ biến hơn của rừng ngẫu nhiên, trong đó các đặc trưng được chọn ngẫu nhiên ở mỗi nút phân chia. Thuật toán rừng ngẫu nhiên có ba siêu tham số chính cần lựa chọn trước khi huấn luyện bao gồm kích thước nút, số lượng cây và số lượng tính năng được lấy mẫu. Trong quần thể rừng ngẫu nhiên, mỗi cây quyết định bao gồm một mẫu dữ liệu được rút ra từ một bộ huấn luyện có sự thay thế, được gọi là mẫu bootstrap. Trong số mẫu huấn luyện đó, một phần ba trong số đó được dành làm dữ liệu thử nghiệm, được gọi là mẫu out-of-bag (oob). Tính ngẫu nhiên được áp dụng thông qua đóng gói các đặc trưng, thêm sự đa dạng hơn cho tập dữ liệu và giảm mối tương quan giữa các cây quyết định. Về hồi quy, các cây quyết định riêng lẻ sẽ được tính trung bình và đối với một nhiệm vụ phân loại, biến phân loại thường gặp sẽ mang vào lớp dự đoán. Cuối cùng, mẫu oob sau đó được sử dụng để xác thực chéo để hoàn thiện dự đoán đó.



Hình 2-2. Mô hình rừng ngẫu nhiên

Lý do đằng sau việc lấy mẫu ngẫu nhiên là một tập hợp các bộ học tập khác nhau có thể bao quát các khía cạnh khác nhau của dữ liệu, bổ sung cho nhau và có hiệu suất cao hơn so với các cây đơn lẻ. Dự đoán cuối cùng của rừng ngẫu nhiên được xác định bởi phiếu bầu đa số của tất cả các cây trong rừng.

2.2.3. Hồi quy Logistic

Hồi quy Logistic là một phương pháp thống kê được sử dụng cho các vấn đề phân loại nhị phân, nơi biến phụ thuộc hoặc kết quả chỉ có thể giả định một trong hai giá trị phân loại. Nó đặc biệt hữu ích trong phân loại chuỗi thời gian, nơi mục tiêu là dự đoán danh mục của một chuỗi quan sát theo thứ tự thời gian.

Mô hình hồi quy logistic được biểu diễn bằng phương trình sau:

$$ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Trong phương trình này, $ln(\frac{\pi}{1-\pi})$ là logit của xác suất của sự kiện (biến phụ thuộc), và X_1, X_2,..., X_k là các biến độc lập. Các $\beta_0, \beta_1, \ldots, \beta_k$ là các tham số của mô hình, được ước lượng từ dữ liệu.

Các tham số beta thường được ước lượng thông qua ước lượng hợp lý tối đa (MLE), kiểm tra các giá trị khác nhau của beta thông qua nhiều lần lặp để tối ưu hóa cho sự phù hợp tốt nhất của log odds.

Kết quả của phân tích hồi quy logistic thường được báo cáo dưới dạng tỷ lệ odds, được thu được bằng cách lũy thừa các ước lượng beta. Tỷ lệ odds đại diện cho tỷ lệ mà một kết quả sẽ xảy ra với một sự kiện cụ thể, so với tỷ lệ của kết quả xảy ra trong sự vắng mặt của sự kiện đó.

Sau khi mô hình đã được tính toán, thì việc thực hành tốt nhất là đánh giá mức độ mô hình dự đoán biến phụ thuộc, được gọi là độ phù hợp. Bài kiểm tra Hosmer–Lemeshow là một phương pháp phổ biến để đánh giá mô hình phù hợp.

Mặc dù hồi quy Logistic là một mô hình tuyến tính, nhưng nó có thể xử lý các mối quan hệ phi tuyến tính giữa các biến đầu vào và đầu ra bằng cách sử dụng các kỹ thuật như thêm các biến đa thức và tương tác (Peng và cộng sự, 2002).

2.2.4. Phân lớp chuỗi thời gian

Phân lớp chuỗi thời gian hay phân tích dữ liệu tuần tự (Time Series Classification - TSC) là một kỹ thuật quan trọng trong học máy, nơi dữ liệu thời gian được phân loại

một cách có hệ thống. Kỹ thuật này liên quan đến việc phân tích chuỗi các điểm dữ liệu được thu thập trong các khoảng thời gian nhất quán. Bản chất của TSC là phân biệt các mẫu phát triển theo thời gian trong dữ liệu, do đó cho phép mô tả các chuỗi quan sát này cho các danh mục cụ thể. Nguyên tắc của dữ liệu chuỗi thời gian các quan sát theo thứ tự thời gian mã hóa các đặc điểm thời gian nội tại, khi được phân tích chính xác, tiết lộ những hiểu biết sâu sắc về các quá trình cơ bản đang diễn ra (Esling và Agon, 2012).

Theo định nghĩa của Theissler và cộng sự (2022), một tập dữ liệu phân loại chuỗi thời gian $D = (X, Y)$ là một tập gồm n chuỗi thời gian, $X = \{x_1, x_2, \dots, x_n\} \in R^{n\,x\,m\,x\,d}$, với một vectơ các nhãn được gán (hoặc các lớp) $Y = \{y_1, y_2, \dots, y_n\} \in N^n$.

Đối với tập dữ liệu D chứa l lớp, yi có thể nhận l giá trị khác nhau. Khi l = 2, D là một tập dữ liệu phân loại nhị phân, trong khi đối với l > 2 , D là một tập dữ liệu phân loại đa lớp. Từ cơ sở này, có thể định nghĩa bài toán phân loại chuỗi thời gian (TSC) như sau:

Cho một tập dữ liệu phân lớp chuỗi thời gian D , phân lớp chuỗi thời gian là nhiệm vụ của việc huấn luyện một hàm hoặc ánh xạ hàm f từ dữ liệu đầu vào X đến một phân phối xác suất cho các lớp ở Y.

Hàm TSC kết quả f nhận đầu vào là một chuỗi thời gian x và trả về nhãn t của lớp mà x thuộc về theo những gì f đã học được, tức là $y = f(x)$. Cách gọi $Y = f(X)$ được sử dụng làm cách gọi tắt cho $Y = \{f(X) \mid x \in X\}$.

Trong những thập kỷ qua, nhiều thuật toán đã được đề xuất để cải thiện hiệu suất dự đoán và khả năng mở rộng của các mô hình tiên tiến. Các cách tiếp cận này bao gồm từ việc tìm ra các số liệu mới, phát triển các mô hình, chuỗi thời gian hình ảnh cho đến mạng lưới thần kinh nhân tạo. Lĩnh vực này có nhiều ứng dụng thực tế và được tìm thấy trong nhiều lĩnh vực khác nhau do tính chất phi cấu trúc của chuỗi thời gian.

2.2.5. MiniRocket

Minimally Random Convolutional Kernel Transform (hay MINIROCKET) là một phép biến đổi dựa trên kernel tích chập ngẫu nhiên tối thiểu. Là phương pháp sử dụng một số lượng nhỏ các kernel tích chập ngẫu nhiên để biến đổi dữ liệu chuỗi thời

gian, MiniRocket giữ lại hai yếu tố quan trọng nhất của Rocket bao gồm: sự thay đổi kích thước của các kernel chập (dilation) và PPV (tức 'proportion of positive values' pooling, một cách để tóm tắt thông tin từ output của một kernel chập, giúp tạo ra một đặc trưng có thể được sử dụng để huấn luyện một bộ phân loại (Tan và cộng sự, 2022). Phương pháp này biến đổi chuỗi thời gian đầu vào sử dụng các kernel chập ngẫu nhiên, và sử dụng các đặc trưng đã biến đổi để huấn luyện một bộ phân loại tuyến tính. MiniRocket tận dụng các tính chất của các kernel, và của PPV, để giảm đáng kể thời gian cần thiết cho việc biến đổi. Các lớp phân loại được sử dụng với MiniRocket giống như Rocket để học cách ánh xạ từ các đặc trưng đến các lớp. Cụ thể, một bộ phân loại có thể có nhiều biến thể tuy nghiên phổ biến thương là hồi quy ridge hoặc hồi quy logistic, được huấn luyện bằng phương pháp Adam (Kingma & Ba, 2014) để tối ưu các phần ngẫu nhiên của mô hình sau đó thực hiện việc dự đoán phân loại.

2.2.6. Shapley Additive Explanations

Shapley Additive Explanations, thường được gọi là SHAP, là phương pháp được sử dụng để giải thích đầu ra của các mô hình máy học. Nó sử dụng giá trị Shapley, cách tiếp cận của lý thuyết trò chơi để đo lường sự đóng góp của mỗi người chơi vào kết quả cuối cùng. Trong đó, mỗi đặc trưng được gán một giá trị quan trọng thể hiện sự đóng góp của nó vào đầu ra của mô hình. Giá trị SHAP cho thấy mỗi đặc trưng ảnh hưởng như thế nào đến từng dự đoán cuối cùng, tầm quan trọng của từng đặc trưng so với các đặc trưng khác và mức độ phụ thuộc của mô hình vào sự tương tác giữa các đặc trưng. Các giá trị SHAP không phụ thuộc vào mô hình, nghĩa là chúng có thể được sử dụng để diễn giải bất kỳ mô hình học máy nào. Nhiều mô hình học máy hiện nay khá phức tạp và khó giải thích được kết quả dự đoán vì vậy SHAP được đề xuất bởi Lundberg và cộng sự (2017) như một giải pháp. Nó giúp giải quyết sự bù trừ giữa độ chính xác và khả năng diễn giải trong các thuật toán phức tạp.

Giải thích giá trị Shapley được biểu diễn dưới dạng phương pháp phân bổ tính năng bổ sung mô hình tuyến tính. Mô hình giải thích SHAP được biểu diễn bởi phương trình:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

trong đó g là mô hình giải thích, $z' \in \{0,1\}^M$ là vectơ liên minh, M là quy mô liên minh

tối đa, và $\phi_i \in$ R là đóng góp của đặc trưng j, giá trị Shapley. Để tính toán các giá trị Shapley, chúng ta mô phỏng rằng chỉ một số giá trị đặc trưng đang "chơi" ("hiện diện") và một số không ("vắng mặt"). Việc biểu diễn dưới dạng mô hình tuyến tính của các liên minh là một cách để tính toán các giá trị ϕ's. Phương trình được tối giản thành:

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_i$$

Giá trị Shapley là giải pháp duy nhất thỏa mãn các đặc tính về Hiệu suất (Efficiency), Đối xứng (Symmetry), Giả (Dummy) và Độ cộng (Additivity). Ngoài ra, SHAP mô tả ba thuộc tính cần có bao gồm Độ chính xác cục bộ (Local accuracy), Sự vắng mặt (Missingness), và Tính nhất quán (Consistency).

Lundberg and Lee (2017) cũng đã đề xuất KernelSHAP, một phương pháp thay thế dựa trên ước lượng kernel cho các giá trị Shapley, lấy cảm hứng từ các mô hình thay thế cục bộ (local surrogate models). Kernel SHAP giúp cải thiện hiệu quả của ước lượng giá trị SHAP không phụ thuộc vào mô hình bằng cách hạn chế sự chú ý vào các loại mô hình cụ thể, đồng thời giúp làm giảm độ phức tạp tính toán và giảm tốc độ thực thi phương pháp.

Kernel SHAP bao gồm 5 bước:

Bước 1: Lấy mẫu các liên minh $z'_k \in \{0,1\}^M, k \in \{1,\ldots,K\}$ (1 = đặc trưng có trong liên kết, 0 = đặc trưng không có trong liên kết).

Bước 2: Nhận dự đoán cho từng z'_k bằng cách trước tiên chuyển đổi z'_k thành không gian đặc trưng ban đầu và sau đó áp dụng mô hình $\hat{f} : \hat{f}(h_x(z'_x))$

Bước 3: Tính trọng số cho mỗi z'_k bằng hạt nhân (kernel).

Bước 4: Huấn luyện mô hình tuyến tính có trọng số.

Bước 5: Trả về giá trị Shapley ϕ_k, tương đương với các hệ số tương quan từ mô hình tuyến tính.

2.2.7. Time Series K-means

Theo nghiên cứu Huang và cộng sự (2016), Time Series K-means (TSKmeans) cho phân cụm dữ liệu chuỗi thời gian được định nghĩa là K-means được kết hợp với

trọng số. Nó cố gắng làm cho khoảng cách giữa các điểm dữ liệu chứa trong một cụm và trung tâm của cụm trở nên nhỏ thông qua việc sử dụng trọng số của nhãn thời gian.

Giả sử cho $X = \{X_1, X_2, \ldots, X_n\}$ là một tập hợp n các chuỗi thời gian. Mỗi mô hình $X_1 = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ là mẫu thứ i được đặc trưng bởi m giá trị, tức là m dấu thời gian. Ma trận thành viên U là ma trận nhị phân n × k, k là tổng số cụm, với $u_{ij} = 1$ chỉ ra rằng X_i thuộc cụm p và $u_{ij}, j \neq p$, là 0. Tâm và trọng số của cụm được biểu thị bằng hai bộ k vectơ $Z = \{Z_1, Z_2, \ldots, Z_k\}$ và $W = \{W_1, W_2, \ldots, W_k\}$, với w_{pj} là trọng số của dấu thời gian thứ j cho cụm thứ p. Mục đích của TSKmeans là tối thiểu hóa hàm mục tiêu sau:

$$P(U, Z, W) = \sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ip} w_{pj} (x_{ij} - z_{ij})^2 + \frac{1}{2}\alpha \sum_{p=1}^{k}\sum_{j=1}^{m-1} (w_{pj} - w_{p,j+1})^2$$

tùy thuộc vào:

$$\begin{cases} \sum_{p=1}^{k} u_{ip} = 1,\ u_{ip} \in \{0, 1\} \\ \sum_{p=1}^{m} w_{pj} = 1, 0 \leq w_{pj} \leq 1 \end{cases}$$

trong đó α là tham số được sử dụng để cân bằng các tác động giữa độ phân tán của các đối tượng trong cụm và độ mịn của trọng số của dấu thời gian. Độ mịn của trọng số giữa các dấu thời gian liền kề tăng lên khi giá trị của α tăng lên. Mục đầu tiên của hàm mục tiêu $P(U, Z, W)$ nhằm mục đích giảm thiểu tổng số lần phân tán của tất cả các cụm. Mục thứ hai của hàm mục tiêu là làm mịn trọng số của các dấu thời gian liền kề. Trong quá trình phân cụm, hàm mục tiêu này đồng thời giảm thiểu sự phân tán bên trong cụm và làm mịn trọng số của dấu thời gian liền kề.

Khoảng cách từ một vectơ Xi của khách hàng i đến trung tâm được tính bằng độ đo DTW. Dynamic time warping (DTW) dựa trên Levenshtein khoảng cách (còn gọi là chỉnh sửa khoảng cách). Nó tìm thấy sự liên kết (hoặc khớp nối) tối ưu giữa hai chuỗi giá trị số và ghi lại những điểm tương đồng linh hoạt bằng cách căn chỉnh tọa độ bên trong cả hai chuỗi. Chi phí của sự liên kết tối ưu có thể được tính toán đệ quy bằng:

$$D(A_i, B_j) = \delta(a_i, b_j) + min \begin{Bmatrix} D(A_{i-1}, B_{j-1}) \\ D(A_i, B_{j-1}) \\ D(A_{i-1}, B_j) \end{Bmatrix}$$

trong đó A_i là dãy con $< a_i, \ldots, a_i >$. Tương đồng tổng thể được đưa ra bởi $D(A_{|A|,} B_{|B|}) = D(A_T, B_T)$.

2.2.8. Các chỉ số đánh giá

Trong quá trình xây dựng mô hình, để xem xét đó có phải một mô hình chất lượng và có thể được đưa vào sử dụng trong thực tế doanh nghiệp hay không, bước đánh giá hiệu suất mô hình là tất yếu. Hiệu quả của mô hình được đánh giá thông qua kết quả quá trình chạy trên tập kiểm thử. Giả sử y_predict là kết quả chạy chạy trên tập kiểm thử và y là giá trị thực tế, ta sẽ so sánh hai giá trị này với nhau. Các chỉ số đánh giá quan trọng được sử dụng phổ biến trong nhiều nghiên cứu bao gồm Accuracy, Precision, Recall, F1 score, Log Loss, AUC.

Độ chính xác là chỉ số được sử dụng phổ biến nhất trong các bài nghiên cứu dự đoán rời bỏ nói chung và nghiên cứu sử dụng phân lớp chuỗi thời gian nói riêng. Độ chính xác được tính bằng tỷ lệ số giá trị dự đoán đúng trên tổng số giá trị của tập dữ liệu.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Tuy vậy, độ chính xác chỉ thể hiện được tỷ lệ phần trăm số giá trị được sắp xếp vào đúng lớp nhưng cụ thể độ chính xác của từng loại dữ liệu. Do đó, Precision, Recall và F1 score được sử dụng:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

$$F1\ score = \frac{2TP}{2\,TP + FN + FP}$$

Các kết quả True Positive, True Negative, False Positive, False Negative sẽ được biểu diễn thông qua bảng ma trận nhầm lẫn – confusion matrix như Bảng *2-1*:

Bảng 2-1. Confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Đồng thời, bài nghiên còn sử dụng hai chỉ số khác là AUC và Log loss.

AUC là thước đo tổng thể độ nhạy $\frac{TP}{TP + FN}$ và độ đặc hiệu $\frac{TN}{TN + FP}$ trên một phạm vi rộng của tất cả các ngưỡng giá trị có thể có. Hiện tại, nó là một thước đo đánh giá hiệu suất được sử dụng rộng rãi phổ biến nhất.

$$\hat{A} = \frac{S_0 - n_o(n_0 + 1)/2}{n_0 n_1}$$

trong đó, n_0 và n_1 đại diện cho số lượng trường hợp tích cực và tiêu cực tương ứng, hơn nữa $S_0 = \sum r_i$, trong đó r_i biểu thị thứ hạng của trường hợp tích cực thứ i trong danh sách được xếp hạng.

Log loss (hay còn gọi là cross-entropy loss) là một thước đo quan trọng được lựa chọn trong việc đánh giá các thuật toán phân loại học máy. Mất log dựa trên xác suất dự đoán trong đó giá trị mất log thấp hơn có nghĩa là dự đoán tốt hơn. Do đó, log-loss rất hữu ích để so sánh các mô hình không chỉ về đầu ra mà còn về kết quả xác suất của chúng. Giả sử cho một vectơ $u = (u_1, \ldots, u_N) \in [0,1]^N$ và nhãn $\sigma \in \{0,1\}^N$, log loss trên vectơ u đối với σ được ký hiệu là $LLoss\ (u;\ \sigma)$ được xác định như sau (Aggarwal và cộng sự, 2021):

$$LLoss\ (u;\sigma) := \frac{-1}{N} ln\left(\prod_{i=1}^{N} u_i^{\sigma_i}(1-u_i)^{1-\sigma_i}\right)$$

2.3. Các nghiên cứu liên quan

Bảng 2-2. Bảng tổng hợp nghiên cứu sử dụng phương pháp dự đoán rời bỏ truyền thống

STT	Tên bài nghiên cứu	Tác giả	Năm	Phương pháp	Kết quả	Đóng góp
1	A Survey on Customer Churn Prediction using Machine Learning Techniques	Saran Kumar A., Chandrakala D.	2016	Nghiên cứu tổng hợp các phương pháp được sử dụng cho quá trình dự đoán của khách hàng.	Nghiên cứu cho thấy cách kết hợp SVM với các thuật toán tăng cường độ chính xác và hiệu suất cao hơn có thể được coi là một công việc trong tương lai để dự đoán.	Đánh giá các thuật toán học máy được sử dụng để dự đoán. Áp dụng các thuật toán này trong các lĩnh vực khác nhau phụ thuộc vào sự tham gia của khách hàng
2	Churn prediction of subscription user for a music streaming service	Sravya Nimmagadda, Akshay Subramaniam, Man Long Wong	2017	Sử dụng XGBoost, Neural Network và hồi quy Logistic với dữ liệu cân bằng và dữ liệu đầy đủ.	Nghiên cứu kết luận rằng thuật toán tăng cường độ dốc hoạt động tốt hơn so với hồi quy logistic và mạng lưới thần kinh để dự đoán tình trạng rời bỏ.	Kỹ thuật trí tuệ nhân tạo rõ ràng đang được sử dụng nhiều hơn để dự đoán báo cáo của khách hàng viễn thông đồng thời cải thiện nhiều hơn.
3	Clustering Prediction Techniques in Defining and Predicting Customers Defection:	Ait Daqud Rachid , Amine Abdellah, Bouikhalene Belaid, Lbibb Rachid	2018	Tích hợp phương pháp k-means và mô hình Length-Recency-Frequency-Monetary (LRFM) để phân cụm khách	Hiệu suất dự đoán của ba mô hình dựa trên xác thực chéo gấp 10 lần. Trung bình, các mô hình dự đoán cung cấp độ chính xác cao hơn 93%. Tập hợp cây quyết định cung cấp kết quả tốt nhất.	Tích hợp được nhiều phương pháp theo các giai đoạn qua đó xác định được các nhóm khách hàng trung thành, rời bỏ hoàn toàn và rời bỏ một nửa. Giúp doanh nghiệp cải

	The Case of E-Commerce Context			hàng, xác định khách hàng rời bỏ, sau đó dự đoán đa lớp dựa trên ba kỹ thuật phân loại: Cây quyết định đơn giản, Mạng nơ-ron nhân tạo và Tập hợp cây quyết định.		thiện việc giữ chân khách hàng. Ngoài ra, còn chứng minh được sự hiệu quả của mô hình tập hợp các cây quyết định.
4	Research Trends in Customer Churn Prediction: A Data Mining Approach	Zhang Tianyuan & Sérgio Moro	2021	Nghiên cứu trình bày một đánh giá tài liệu về dự đoán của khách hàng dựa trên 40 bài báo có liên quan được xuất bản từ năm 2010 đến tháng 6 năm 2020.	Nghiên cứu đã chứng minh rằng các kỹ thuật khai thác dữ liệu được sử dụng rộng rãi nhất là cây quyết định (DT), máy vectơ hỗ trợ (SVM) và hồi quy logistic (LR).	Khám phá và so sánh các phương pháp khai thác dữ liệu và học máy khác nhau để dự đoán tỷ lệ rời bỏ khách hàng trong ngành viễn thông và ngành dịch vụ tài chính.
5	A survey on machine learning methods for churn prediction	Louis Geiler, Séverine Affeldt & Mohamed Nadif	2022	Nghiên cứu hành vi của 11 phương pháp học có giám sát và bán giám sát cùng với bảy phương pháp lấy mẫu trên 16 tập dữ liệu khác nhau và công khai về khách	Các thử nghiệm dẫn đến đề xuất thực tế cho quy trình dự đoán dựa trên cách tiếp cận tổng hợp.	Nghiên cứu rút ra các hướng dẫn chung từ một chuẩn mực của các kỹ thuật học máy được giám sát kết hợp với các phương pháp lấy mẫu dữ liệu được sử dụng rộng rãi trên các bộ dữ liệu có sẵn công khai trong bối cảnh dự

			hàng rời bỏ.		đoán.
6	Research on customer churn prediction and model interpretability analysis	Ke Peng,Yan Peng, Wenguang Li	2023	Sử dụng nhiều phương pháp lấy mẫu để cân bằng dữ liệu và xây dựng mô hình dự đoán báo cáo của khách hàng ngân hàng để nhận dạng churn bằng GA-XGBOOST.	SMOTEENN được áp dụng hiệu quả hơn SMOTE và ADASYN trong việc xử lý tình trạng mất cân bằng dữ liệu ngân hàng. Giá trị F1 và AUC của mô hình được XGBoost cải thiện và tối ưu hóa bằng thuật toán di truyền có thể đạt lần lượt 90% và 99%, tối ưu so với sáu mô hình học máy khác.	Nghiên cứu có thể cung cấp thông tin hữu ích từ mô hình blackbox dựa trên việc xác định chính xác khách hàng bị khuất phục, có thể cung cấp tài liệu tham khảo cho các ngân hàng thương mại để cải thiện chất lượng dịch vụ của họ và giữ chân khách hàng.
7	Customer Churn in Subscription Business Model— Predictive Analytics on Customer Churn	Boyuan Zhang	2023	Ba thuật toán, cụ thể là hồi quy logistic, tăng độ dốc (SMOTE) và mạng lưới thần kinh, vượt trội so với 6 thuật toán khác về mặt dự đoán tỷ lệ rời bỏ của khách hàng.	Bài có so sánh hiệu suất của chín thuật toán khác nhau và xác định thuật toán hoạt động tốt nhất, hồi quy logistic, với độ chính xác dự đoán là 79,6%. Việc sử dụng Smote trong quá trình xây dựng mô hình và gợi ý rằng việc lấy mẫu hoặc lấy mẫu dữ liệu có thể hiệu quả hơn trong việc cân bằng các bộ dữ liệu.	Cung cấp thông tin chi tiết cho các doanh nghiệp đăng ký về việc chọn thuật toán hiệu quả để dự đoán khu vực khách hàng và thực hiện các thay đổi chủ động để giảm tỷ lệ khuấy. Sức mạnh nằm trong phân tích kỹ lưỡng và xác định thuật toán hoạt động tốt nhất, hồi quy logistic. Tuy nhiên, điểm yếu có thể là cuộc thảo luận hạn chế về những hạn chế của nghiên cứu và tính tổng quát của các phát hiện.

Bảng 2-3. Bảng tổng hợp nghiên cứu ứng dụng phân lớp chuỗi thời gian và phương ph giải thích mô hình phân lớp chuỗi thời gianáp

STT	Tên bài nghiên cứu	Tác giả	Năm	Phương pháp	Kết quả	Đóng góp
1	Time Series Classification	M. M. Gabr, L. M. Fatehy	2013	Bài viết đề xuất một phiên bản sửa đổi của phân tích phân biệt cổ điển sử dụng phân tích thành phần chính (PCA) để khắc phục tính chất đa chiều của dữ liệu chuỗi thời gian	Phương pháp phân loại đã được sửa đổi bằng cách sử dụng phân tích thành phần chính (PCA) cho thấy tỷ lệ phân loại chính xác, đặc biệt là khi so sánh với các phương pháp khác. Phương pháp sửa đổi PCA cũng được áp dụng cho một bộ dữ liệu chuỗi thời gian thực và chứng minh tỉ lệ phân loại chính xác vượt trội.	Những phát hiện của nghiên cứu này góp phần vào lĩnh vực khai thác dữ liệu và cung cấp một cách tiếp cận có giá trị để phân loại dữ liệu chuỗi thời gian trong các ứng dụng khác nhau.
2	Highly comparative feature-based time-series classification	Ben D. Fulcher, Nick S. Jones	2014	Sử dụng lựa chọn đặc trưng tiến lên tham lam với một bộ phân loại tuyến tính để chọn ra những đặc trưng thông tin nhất về cấu trúc lớp từ hàng nghìn đặc trưng đã tính toán cho mỗi chuỗi thời gian trong tập dữ liệu huấn luyện.	Phương pháp này cho phép giảm bậc của chiều không gian đáng kể, cho phép nó hoạt động tốt trên các bộ dữ liệu rất lớn chứa chuỗi thời gian dài hoặc chuỗi thời gian có độ dài khác nhau. Hiệu suất phân loại vượt trội so với các bộ phân loại dựa trên thực thể thông thường, bao gồm bộ phân loại một láng giềng gần nhất sử dụng khoảng cách Euclid và biến đổi thời gian động	Nghiên cứu cung cấp một phương pháp mới cho phân loại chuỗi thời gian, giúp đơn giản hóa quy trình lựa chọn tính năng và cải thiện độ chính xác phân loại.
3	Scalable time series	Patrick Schäfer		Mô hình BOSS VS mở rộng từ mô hình BOSS	Mô hình đề xuất được chứng minh là chính xác, nhanh chóng và ít bị ảnh	Mô hình đề xuất được chứng minh là chính

	classification			bằng cách biểu diễn nhỏ gọn các lớp và sử dụng ma trận tf-idf để phân loại.	hưởng bởi nhiễu.	xác, nhanh chóng và ít bị ảnh hưởng bởi nhiễu.
4	A review on distance based time series classification	Amaia Abanda, Amaia Abanda, Usue Mori, Jose A. Lozano	2018	Tổng hợp và thảo luận các cách tiếp cận khác nhau trong việc phân loại chuỗi thời gian dựa trên khoảng cách, tiêu biểu là phương pháp 1-NN	Phương pháp 1-NN (nearest neighbor) đã được sử dụng rộng rãi trong phân loại chuỗi thời gian dựa trên khoảng cách vì tính đơn giản nhưng vẫn hiệu quả. Tuy nhiên, hiệu suất cao của nó có thể được quy cho việc sử dụng các khoảng cách cụ thể cho chuỗi thời gian trong quá trình phân loại, chứ không phải do chính bản thân bộ phân loại. Các phương pháp mới đã xuất hiện trong những năm gần đây và chúng cạnh tranh hoặc vượt trội hơn so với các phương pháp dựa trên 1-NN.	Nghiên cứu đã đưa ra một cái nhìn tổng quan về các phương pháp phân loại chuỗi thời gian dựa trên khoảng cách.
5	Time series for early churn detection: Using similarity based classification for dynamic networks	María Óskarsdóttir, Tine Van Calster, Bart Baesens, Wilfried Lemahieu, Jan Vanthienen	2018	Phương pháp rừng tương tự dựa trên khoảng cách giữa các chuỗi thời gian và được sử dụng kết hợp với rừng ngẫu nhiên.	Phương pháp đề xuất sử dụng dữ liệu chuỗi thời gian được trích xuất từ các mạng cuộc gọi để thể hiện hành vi động của khách hàng đã cho thấy kết quả đầy hứa hẹn trong việc dự đoán tình trạng gián đoạn trong ngành viễn thông. Phương pháp rừng tương tự, cùng với các phần mở rộng được đề xuất, vượt trội hơn các phương pháp dự đoán tỷ lệ rời bỏ truyền	Mở rộng phương pháp rừng tương tự để phù hợp với chuỗi thời gian đa biến, cung cấp thông tin chi tiết về kỹ thuật phân loại để phân tích hành vi của khách hàng.

					thống và hoạt động tốt trong việc dự đoán tỷ lệ rời bỏ sớm.	
6	Deep learning for time series classification: a review	Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar,Pierre-Alain Muller	2019	Nghiên cứu đánh giá hiệu suất của các kiến trúc DNN khác nhau, bao gồm Multi Layer Perceptron (MLP), Mạng thần kinh chuyển đổi (CNN) và Mạng trạng thái Echo (ESN) dựa trên tiêu chuẩn TSC đơn biến (kho lưu trữ UCR/UEA) và 12 bộ dữ liệu chuỗi thời gian đa biến.	ResNet vượt trội đáng kể so với các mô hình khác, với độ chính xác cao nhất, 50 trên 85 bộ dữ liệu và vượt trội đáng kể so với kiến trúc FCN. Tuy nhiên, ResNet, PF, ST và BOSS thể hiện hiệu suất tương tự theo các thử nghiệm thống kê, ResNet không có sự khác biệt đáng kể so với COTE cho thấy rằng việc đánh giá sâu hơn trên nhiều bộ dữ liệu hơn sẽ cung cấp những hiểu biết sâu sắc hơn về hiệu suất của chúng.	Một khung học sâu nguồn mở đã được cung cấp cho cộng đồng TSC, tạo điều kiện thuận lợi cho việc nghiên cứu và thử nghiệm sâu hơn trong lĩnh vực này.
7	Dynamic behavior based churn prediction in mobile telecom	Nadia Alboukaey, Ammar Joukhada, Nada Ghneim	2020	Đề xuất mô hình phân loại chuỗi thời gian dự đoán rời bỏ theo ngày, dựa trên các mô hình RFM-based, LSTM-based và CNN-based.	Phương pháp dự đoán rời bỏ bằng bằng phân lớp chuỗi thời gian cho phép phát hiện sớm quyết định rời bỏ của khách hàng. Các chỉ số đo đo lường cho kết quả không cao nhưng cho thấy được tính ứng dụng của phương pháp này.	Đề xuất phương phương pháp dự đoán rời bỏ có tính ứng dụng cao dựa trên phân lớp chuỗi thời gian.
8	Fast and Accurate Time Series Classification Through Supervised	Nestor Cabello, Elham Naghizade, Jianzhong Qi, Lars Kulik	2020	phương pháp Randomized-Supervised Time Series Forest (r-STSF), một cách tiếp cận dựa trên khoảng thời gian hiệu quả để phân	Kết quả thử nghiệm cho thấy r-STSF đạt được độ chính xác hàng đầu và tốc độ nhanh hơn nhiều so với hầu hết các phương pháp phân loại chuỗi thời gian (TSC) tại thời điểm đó.	Đối với phương pháp phân lớp dựa trên tần số và khoảng thời gian, nghiên cứu đã chứng minh r-STSF là một phương pháp hiệu

	Interval Search			loại chuỗi thời gian dựa trên giá trị tổng hợp của các chuỗi phụ có tính phân biệt (khoảng thời gian)(Interval- and Frequency-Based Methods).		quả.
9	ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels	Angus Dempster,François Petitjean, Geoffrey I.Webb	2020	Phân loại tuyến tính đơn giản với Rocket(các hạt nhân ngẫu nhiên)	Các phân loại tuyến tính đơn giản bằng cách sử dụng các hạt nhân ngẫu nhiên đạt được độ chính xác tiên tiến.	Phương pháp phân loại tuyến tính đơn giản sử dụng các hạt nhân ngẫu nhiên đạt được độ chính xác hiện đại với một phần chi phí tính toán của các phương pháp hiện có để phân loại chuỗi thời gian.
10	MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification	Angus Dempster Daniel F. Schmidt Geoffrey I. Webb	2021	Phân loại tuyến tính đơn giản với Rocket(các hạt nhân ngẫu nhiên) tuy nhiên giảm bớt các kernel ngẫu nhiên	MiniRocket đạt được độ chính xác cao nhất để phân loại chuỗi thời gian. MiniRocket nhanh hơn Rocket tới 75 lần trên các tập dữ liệu lớn hơn.	MiniRocket nhanh hơn nhiều so với bất kỳ phương pháp chính xác tương đương nào khác (bao gồm cả Rocket) và chính xác hơn nhiều so với bất kỳ phương pháp nào khác có chi phí tính toán tương tự.
11	Survey of early	Mengchen	2021	Khảo sát về các phương	Kết quả nghiên cứu nhấn mạnh được	Nghiên cứu giới thiệu

	time series classification methods	YANG, Xudong CHEN, Peng CAI, Lyu NI		pháp phân loại chuỗi thời gian sớm, bao gồm các phương pháp và ứng dụng khác nhau trong lĩnh vực này	tầm quan trọng của việc phân loại dữ liệu chuỗi thời gian sớm cho các ứng dụng như dự báo chỉ số chứng khoán, lập kế hoạch quy trình, phân loại ECG và hệ thống vận chuyển.	các kỹ thuật sáng tạo để phân loại sớm dữ liệu chuỗi thời gian, có thể có các ứng dụng trong các lĩnh vực khác nhau như chăm sóc sức khỏe, IoT và nhận dạng hoạt động của con người.
12	Hybrid Model for the Customer Churn Prediction	Mansimar Anand, Irtibat Shaukat, Harnoor Kaler, Jai Narula, Prashant Singh Rana	2021	Sử dụng phương pháp phân cụm K-means để tạo ra các cụm, sau đó áp dụng thuật toán phân loại Cây quyết định để dự đoán mô hình. Mô hình đã được thử nghiệm với bảy thuật toán: Cây quyết định, Hồi quy Logistic, Naive Bayes, K-nearest neighbors, Gradient Boosting, Ada Boosting, và K-means	Mô hình được đề xuất đã đạt được độ chính xác 91.48% khi được xác thực bằng phương pháp cross-validation K-fold	Nghiên cứu đã mang đến một mô hình dự đoán tốt cho doanh nghiệp, kết hợp bảy mô hình riêng lẻ, bao gồm rừng ngẫu nhiên, hồi quy logistic, hàng xóm k-gần nhất, Bayes ngây thơ, tăng cường ADA, tăng độ dốc và K-MEAN, để tạo ra một mô hình lai mới đạt được độ chính xác cao.
13	Explainable AI for time series classification: a review, taxonomy and	Andreas Theissler, Francesco Spinnato, Udo Schlegel,	2022	Trình bày bài đánh giá tài liệu sâu rộng đầu tiên về AI có thể giải thích (XAI) để phân loại chuỗi thời gian, phân loại lĩnh	Các phương pháp được phân loại thành ba loại: giải thích dựa trên điểm thời gian, dựa trên chuỗi con và dựa trên trường hợp.	Cung cấp cái nhìn tổng quan về Explainable AI trong phân lớp chuỗi thời gian.

	research directions	Riccardo Guidotti		vực nghiên cứu thông qua phân loại. Các tác giả xác định các hướng nghiên cứu mở về loại hình giải thích và đánh giá khả năng diễn giải.		
14	A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications	Will Ke Wang , Ina Chen, Leeor Hershkovich, Jiamu Yang, Ayush Shetty, et al,.	2022	Tổng hợp các bài viết về các kỹ thuật phân loại chuỗi thời gian được sử dụng trong các ứng dụng y sinh	Các mô hình học máy như Cây quyết định, CNN, SVM và kNN đã được sử dụng cho các nhiệm vụ như nhận dạng bước đi, nhận dạng cá nhân bằng EMG, dự đoán khả năng phục hồi sau chấn thương tủy sống và phân loại giấc ngủ và thức từ dữ liệu PPG, đạt được điểm số có độ chính xác cao từ 0,9523 đến 0,9981.	Những phát hiện này đóng góp những hiểu biết có giá trị về tính hiệu quả của các thuật toán này trong các ứng dụng theo dõi sức khỏe.
15	Time Series Classification with Shapelet and Canonical Features	Hai-Yang Liu, Zhen-Zhuo Gao, Zhi-Hai Wang, Yun-Hao Deng	2022	Tích hợp nhiều đặc trưng chuỗi thời gian chuẩn vào Shapelets để tăng khả năng thích ứng với các vấn đề phân loại khác nhau và bù đắp cho sự mất mát độ chính xác do việc chọn lọc ngẫu nhiên.	Kết quả thực nghiệm trên 112 bộ dữ liệu chuỗi thời gian UCR cho thấy thuật toán SCF chính xác hơn thuật toán STC dựa trên tìm kiếm Shapelet chính xác và kỹ thuật biến đổi Shapelet, cũng như nhiều thuật toán phân loại chuỗi thời gian tiên tiến khác.	Bài báo đề xuất một phương pháp mới giúp giảm thiểu chi phí tính toán cao thường gặp trong các phương pháp dựa trên Shapelets
16	Agnostic Local Explanation for	Maël Guillemé; Véronique	2019	LEFTIST là một phương pháp giải thích cục bộ	Mô hình phân lớp chuỗi thời gian ResNet đã đạt được hiệu suất tương	Phân tách các khung giải thích hiện có

	Time Series Classification	Masson; Laurence Rozé; Alexandre Termier		cho các mô hình phân loại chuỗi thời gian bộ bất khả tr, cung cấp giải thích cho các dự đoán được đưa ra bởi bất kỳ mô hình phân loại chuỗi thời gian nào.	tự như COTE, một trong những mô hình phân lớp tốt nhất. Ngoài ra, nghiên cứu cũng đã tiến hành một cuộc khảo sát người dùng để hiểu rõ hơn về cách các phân loại dự đoán. Kết quả cho thấy, trong những trường hợp dễ dàng (phân loại không mắc lỗi), giải thích của LEFTIST giúp người dùng hiểu được dự đoán của mô hình phân lớp.	thành các thành phần cơ bản, chỉ ra cách chúng có thể được điều chỉnh cho dữ liệu chuỗi thời gian và tái tổ chức, tạo ra một cách tiếp cận linh hoạt và có thể tùy chỉnh. Nhờ vậy, đây là nghiên cứu thực nghiệm đầu tiên về giải thích cục bộ trong ngữ cảnh của dữ liệu chuỗi thời gian.
17	How can I explain this to you? An empirical study of deep neural network explanation methods	J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia and M. Srivastava	2020	Phương pháp được sử dụng trong nghiên cứu này liên quan đến việc tạo ra giải thích cho một mô hình mạng nơ-ron sâu (DNN).	Nghiên cứu này đã khảo sát hàng trăm người tham gia và kết luận rằng phương pháp giải thích bằng ví dụ và LIME là những phong cách giải thích được ưa chuộng hiện nay theo người dùng cuối không chuyên. Trong các lĩnh vực dữ liệu bao gồm hình ảnh, âm thanh và cảm biến, việc giải thích bằng các ví dụ đào tạo gần nhất cung cấp cho người dùng cơ hội để so sánh các đặc trưng trên một đầu vào kiểm tra và các ví dụ đúng tương tự. Trong lĩnh vực văn bản, phương pháp của LIME trong việc	Nghiên cứu này đã đóng góp bằng cách cung cấp một sự thống nhất, so sánh và phân tích các phương pháp giải thích hiện tại cho DNNs trong nhiều ứng dụng và lĩnh vực đầu vào.

				phân rã và chú thích các đầu vào kiểm tra cung cấp một cách tiếp cận trực quan cho phân loại văn bản.	
18	Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions	Andreas Theissler; Francesco Spinnato; Udo Schlegel; Riccardo Guidotti	2022	Quan sát sự tiến bộ trong việc phát triển các phương pháp giải thích trong phân loại chuỗi thời gian nhằm chứng minh sự cần thiết phải cấu trúc và xem xét lại lĩnh vực này.	Các lĩnh vực nghiên cứu được phân loại bằng cách chia nhỏ các phương pháp thành dựa trên thời gian, dựa trên chuỗi con và dựa trên trường hợp.	Trình bày thành công bài đánh giá đầu tiên về (XAI) để giải thích phân loại chuỗi thời gian. Đồng thời cũng đã xác định các hướng nghiên cứu tiềm năng liên quan đến phương pháp giải thích và đánh giá các giải thích, khả năng diễn giải.
19	Conceptual challenges for interpretable machine learning	D. S. Watson	2022	Sử dụng Greedy function approximation: A gradient boosting machine, một thuật toán máy học mạnh mẽ giúp xây dựng và tối ưu hóa mô hình dự đoán. Đồng thời, LIME và SHAP được sử dụng để giải thích mô hình máy học, giúp hiểu rõ hơn về dự đoán của mô hình đầu vào.	Nghiên cứu đã chỉ ra rằng hàm phụ thuộc từng phần của Friedman có cấu trúc tương tự như điều chỉnh cửa sau của Pearl khi các biến điều kiện đáp ứng một số điều kiện nhất định. Nói cách khác, sự phụ thuộc từng phần của kết quả đối với một tập hợp tính năng nhất định trong một mô hình dự đoán thuần túy có thể được giải thích tự nhiên như là hiệu ứng nguyên nhân của các biến đó đối với kết quả.	Chỉ ra rằng hàm phụ thuộc từng phần của kết quả đối với một tập hợp tính năng nhất định trong một mô hình dự đoán thuần túy có thể được giải thích tự nhiên như là hiệu ứng nguyên nhân của các biến đó đối với kết quả.

20	A Dictionary-Based Approach to Time Series Ordinal Classification	Rafael Ayllón-Gavilán, David Guijo-Rubio, Pedro Antonio Gutiérrez & César Hervás-Martínez	2023	O-TDE là một sự điều chỉnh của thuật toán TDE hiện tại, được thiết kế để xử lý các vấn đề phân loại chuỗi thời gian có thứ tự (TSOC). Nghiên cứu đã thực hiện một so sánh toàn diện sử dụng một tập hợp 18 vấn đề TSOC để đánh giá hiệu suất của O-TDE so với các kỹ thuật dựa trên từ điển danh nghĩa khác.	Các thí nghiệm cho thấy O-TDE đạt được cải thiện đáng kể so với bốn kỹ thuật dựa trên từ điển danh nghĩa khác, chứng tỏ hiệu quả của việc sử dụng tính chất thứ tự trong phân loại chuỗi thời gian. Nhưng số lượng lớn các mẫu thu được từ phương pháp này có thể lớn, điều này có thể hạn chế khả năng diễn giải của mô hình.	Nghiên cứu mở rộng lĩnh vực phân loại chuỗi thời gian bằng cách giới thiệu khái niệm về TSOC và cung cấp một thuật toán mới cho việc này. O-TDE có thể cải thiện hiệu suất phân loại trong các vấn đề mà nhãn của chuỗi thời gian có một trật tự tự nhiên.
21	Early Churn Prediction from Large Scale User-Product Interaction Time Series	Shamik Bhattacharjee, Utkarsh Thukral, Nilesh Patil	2023	Các phương pháp được thảo luận bao gồm kỹ thuật tính năng, tập hợp thời gian và các mô hình học sâu như Transformer, LSTM.	Mô hình có hiệu suất cao nhất là kiến trúc Transformer, vượt trội hơn dự đoán của các mô hình cổ điển trung bình khoảng 6%. Với cách tiếp cận cải tiến này, có phạm vi đáng kể để khám phá các ứng dụng seq-2-seq trong các ứng dụng chuỗi thời gian khác (ví dụ: dự đoán CLTV, Dự báo, đề xuất sản phẩm tuần tự là một số trường hợp sử dụng hàng đầu). Hạn chế chính của phương pháp này là yêu cầu khối lượng dữ liệu huấn luyện lớn và thời gian huấn luyện dài.	Nghiên cứu cho thấy cách tiếp cận coi dự đoán của Churn là phân loại chuỗi thời gian đa biến, chứng minh rằng việc kết hợp hoạt động của người dùng và mạng lưới thần kinh sâu mang lại kết quả đáng chú ý để dự đoán trong bối cảnh doanh nghiệp đến khách hàng phức tạp.

Trên góc độ của các phương pháp truyền thống nhằm dự đoán rời bỏ, các kỹ thuật học máy hoạt động trên dữ liệu tĩnh đã có những đóng góp đáng kể vào lĩnh vực này. Bài nghiên cứu của Saran Kumar và Chandrakala (2016) đã chỉ ra tiềm năng của SVM kết hợp với các thuật toán tăng cường để cải thiện độ chính xác dự đoán khách hàng rời bỏ. Bên cạnh đó, nghiên cứu của Rachid và cộng sự (2018) đã tích hợp phương pháp phân cụm và các mô hình phân loại để dự đoán rời bỏ khách hàng trong lĩnh vực thương mại điện tử, đem lại độ chính xác cao và hỗ trợ doanh nghiệp cải thiện chiến lược giữ chân khách hàng. Nghiên cứu của Tianyuan và Moro (2021) đã đưa ra một so sánh tổng quan về các xu hướng dự đoán rời bỏ khách hàng bằng các kỹ thuật phổ biến như cây quyết định (Decision Tree) , SVM, và hồi quy logistic, giúp hiểu rõ hơn về các phương pháp đang được áp dụng trong lĩnh vực này. Bằng hướng tiếp cận tương tự, nghiên cứu của Geiler và cộng sự (2022) cũng đã trình bày một bản khảo sát toàn diện về các phương pháp học máy dự đoán rời bỏ, đưa ra các hướng dẫn về cách tiếp cận và áp dụng thực tiễn cho các doanh nghiệp. Các công trình nghiên cứu khác như của Peng và cộng sự (2023) đã tập trung vào việc áp dụng các phương pháp học máy cụ thể như GA-XGBoost và các thuật toán khác để dự đoán khách hàng rời bỏ, đem lại hiệu quả cao và sự hiểu biết sâu sắc về các yếu tố ảnh hưởng đến sự ra đi của khách hàng trong ngành dịch vụ. Tuy nhiên, các mô hình này tiếp cập theo cách truyền thống, tức sử dụng dữ liệu tĩnh, không thể hiện được sự biến đổi của hành vi khách hàng theo thời gian. Pustokhina và cộng sự (2021) đã đưa ra luận điểm rằng việc sử dụng dữ liệu tĩnh sẽ bỏ qua các xu hướng mới nhất trong hành vi khách hàng và việc thực hiện dự đoán theo cấp độ mỗi tháng là không đủ để phản ứng kịp thời với quyết định của khách hàng nếu quyết định rời bỏ được đưa ra vào đầu tháng.

Nghiên cứu về phân loại chuỗi thời gian đã có tác động đáng kể đến lĩnh vực khoa học dữ liệu và ứng dụng trong thực tế. Ismail Fawaz và cộng sự (2019) đã nghiên cứu về hiệu suất của các kiến trúc mạng nơ-ron sâu (DNN) như MLP, CNN và ESN, mở ra những hướng nghiên cứu mới trong phân loại biến chuỗi thời gian đa biến. Một nghiên cứu của Cabello và cộng sự (2020) đã giới thiệu phương pháp r-STSF, một phương pháp phân loại chuỗi thời gian dựa trên khoảng thời gian, mang lại độ chính xác cao và nhanh hơn nhiều so với các phương pháp khác. Nghiên cứu của Theissler và cộng sự (2022) về trí tuệ nhân tạo khai báo cung cấp cái nhìn tổng quan về trí tuệ nhân

tạo giải thích được trong phân loại chuỗi thời gian giúp hướng dẫn nghiên cứu về khả năng diễn giải của kết quả phân loại. Các nghiên cứu về ứng dụng y sinh (Wang và cộng sự, 2022) đã cho thấy khả năng ứng dụng tuyệt vời của việc phân loại chuỗi thời gian, chẳng hạn như trong việc dự đoán khả năng phục hồi sau chấn thương và theo dõi sức khỏe. Các thuật toán mới như Shapelets (Liu và cộng sự, 2022) đã cải thiện hiệu suất phân loại của chuỗi thời gian có thứ tự, góp phần phát triển các phương pháp tiên tiến nhất trong lĩnh vực này. Nghiên cứu của Óskarsdóttir và cộng sự (2018) đã trình bày phương pháp sử dụng rừng tương tự để dự đoán sự gián đoạn trong ngành viễn thông, mang lại kết quả đầy hứa hẹn cho việc ứng dụng chuỗi thời gian trong lĩnh vực viễn thông, đồng thời cho thấy tính ứng dụng cao của mô hình phân lớp sử dụng chuỗi thời gian theo tuần. Hạn chế của nghiên cứu này là việc lấy mẫu thực nghiệm trên cùng một khung thời gian chưa bám sát thực tế triển khai và tính chất của chuỗi thời gian, vì chuỗi thời gian ở hai khung thời gian sẽ có tính chất khác nhau nên mô hình xây dựng được sẽ không đủ tính tổng quát (generalization).

Việc dự đoán rời bỏ dựa vào phân lớp chuỗi thời gian tồn tại khó khăn về độ phức tạp tính toán, điển hình là phương pháp 1-NN DTW có độ phức tạp thời gian bậc hai theo độ dài chuỗi thời gian và độ chính xác của nó bị suy giảm khi có nhiễu (Schäfer, 2016) . Giải pháp cho vấn đề này là một mô hình có khả năng dự đoán chính xác đồng thời không mất nhiều thời gian để huấn luyện, tối thiểu được độ phức tạp không - thời gian (time-space complexity). Dựa vào những phương pháp từ các nghiên cứu khảo lược có thể thấy phương pháp phân lớp dựa trên hạt nhân (kernel), cụ thể là phương pháp MiniRocket được đề xuất bởi Dempster và cộng sự (2020), là phương pháp tối ưu và phù hợp cho phân lớp chuỗi thời gian khi đặt ra yêu cầu về thời gian. Phương pháp này có tốc độ nhanh, gần như mang tính quyết định và về cơ bản duy trì độ chính xác như nhau. MiniRocket nhanh hơn nhiều so với bất kỳ phương pháp chính xác tương đương nào khác (bao gồm cả Rocket, nhanh hơn đến 75 lần) và chính xác hơn nhiều so với bất kỳ phương pháp nào khác có chi phí tính toán tương đương.

Một vấn đề khác đối với các mô hình phân lớp chuỗi thời gian là chọn đặc trưng và khả năng giải thích mô hình. Theo Vankatest và cộng sự (2019), khảo sát cho thấy phần lớn các phương pháp chọn đặc trưng được thiết kế cho dữ liệu tĩnh. Các mô hình

phân lớp chuỗi thời gian giải thích được khá ít và chia ra thành ba nhóm, gồm dựa trên điểm thời gian, dựa trên chuỗi con và dựa trên trường hợp (Theissler và cộng sự, 2022). Guillemé và cộng sự (2019) đã trình bày phương pháp LEFTIST như một cách giải thích cục bộ cho các mô hình phân loại chuỗi thời gian, thành công trong việc phân tách các khung giải thích hiện có thành các thành phần cơ bản, tạo ra một cách tiếp cận linh hoạt và có thể tùy chỉnh. Jeyakumar và cộng sự (2020) đã tiến hành một nghiên cứu điều tra nhằm tạo ra các giải thích cho các mô hình mạng nơ-ron sâu (DNN) và đưa ra hai kết luận có quan trọng: (1) Các phương pháp giải thích dựa trên ví dụ và LIME đang được ưa chuộng trong các lĩnh vực như hình ảnh, âm thanh và văn bản và (2) Khi áp dụng cho chuỗi thời gian, các đóng góp của đặc trưng và phương tiện giải thích như bản đồ nhiệt, thường được sử dụng bởi các chuyên gia trong lĩnh vực nhưng không hiệu quả đối với người dùng thông thường, vì khó diễn giải điểm liên quan nếu không có kiến thức về lĩnh vực. Giải pháp cho vấn đề giải thích mô hình phân lớp chuỗi thời gian cũng được đề cập bởi nghiên cứu của Watson (2022), đã nêu các thách thức khái niệm trong việc làm cho máy học có thể giải thích, sử dụng thuật toán Greedy function approximation cùng với các phương pháp LIME và SHAP để giải thích các mô hình máy học dựa trên các đặc trưng ảnh hưởng.

Lược khảo cho thấy, đối với bài toán phân lớp chuỗi thời gian, MiniRocket là phương pháp có độ chính xác và tốc độ cao nhất. Đồng thời, SHAP và LIME là các phương pháp giải thích mô hình hiệu quả và phổ biến nhất. Do đó, bài nghiên cứu này sẽ tập trung vào việc khám phá một phương pháp tiếp cận mới, linh hoạt và giúp dự đoán rời bỏ sớm hơn, bằng cách xây dựng mô hình MiniRocket-SHAP. Đây là mô hình kết hợp giữa MiniRocket và SHAP, mà qua khảo sát đã cho thấy là rất hiệu quả cho việc phân lớp dữ liệu chuỗi thời gian và giải thích mô hình. Mô hình này hướng tới mục tiêu đảm bảo tốc độ và sự chính xác để dự đoán khách hàng rời bỏ dựa trên chuỗi thời gian và còn có thể giải thích rõ hơn các yếu tố gây rời bỏ và thời điểm ảnh hưởng đến quyết định rời bỏ. Qua đó, nó mở ra một phương pháp tiếp cận mới với nhiều tiềm năng trong bài toán dự đoán rời bỏ.

Tóm tắt chương 2

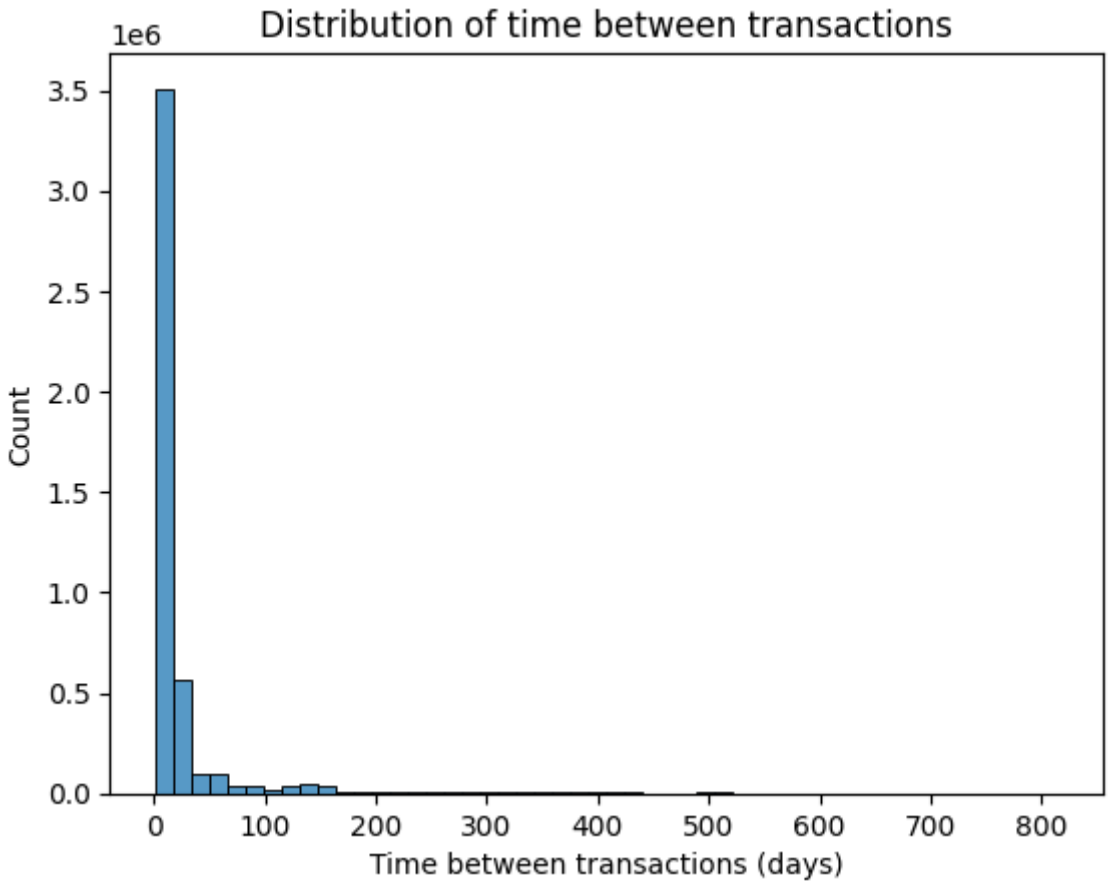
Trong chương 2, nhóm đưa ra một số định nghĩa liên quan đến nghiên cứu như

khái niệm về khách hàng rời bỏ và dự đoán khách hàng rời bỏ. Những kỹ thuật, phương pháp hay chỉ số đánh giá đều được thể hiện rõ trong phần này. Cuối cùng, nhóm thực hiện khảo lược một số nghiên cứu sử dụng phân lớp chuỗi thời gian, đánh giá khoảng trống của chúng và định hướng cải thiện và phát huy điểm mới trong nghiên cứu này.

CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Thiết lập thực nghiệm

3.1.1. Định nghĩa khách hàng rời bỏ



Hình 3-1. Phân phối thời gian giữa các giao dịch

Từ biểu đồ thể hiện phân phối của số ngày giữa các giao dịch được thực hiện, nhóm nghiên cứu thấy rằng 84.67% phân phối nằm trong khoảng từ 0 đến 30 ngày, những khoảng còn lại có mật độ rất thấp. Vì vậy, trong bài toán đặt ra, nhóm nghiên cứu thiết lập định nghĩa khách hàng rời bỏ khi khách hàng không gia hạn dịch vụ sau 4 tuần, tương đương 28 ngày tính từ thời điểm gói đăng ký cuối cùng trong khung thời gian thực nghiệm hết hạn. Trong phạm vi bài nghiên cứu này, các khách hàng rời bỏ được gọi tên là nhóm "Churn" và nhóm khách hàng ở lại được gọi tên là nhóm "Non-churn".

3.1.2. Dữ liệu đầu vào và ma trận chuỗi thời gian đa biến

Dữ liệu về khách hàng có được từ KKBOX (Addison Howard và cộng sự, 2017) bao gồm ba nhóm đặc trưng: hành vi nghe nhạc, thông tin khách hàng và lịch sử giao dịch.

- Hành vi nghe nhạc: dữ liệu tổng hợp mỗi ngày từ hành vi của khách hàng.
- Lịch sử giao dịch: dữ liệu về các giao dịch đăng ký gói, gia hạn, hủy gói mà khách hàng đã thực hiện.
- Thông tin khách hàng: dữ liệu ghi nhận từ khách hàng khi thực hiện đăng ký, gồm thông tin đăng ký và thông tin nhân khẩu học, không thay đổi theo thời gian.

Trong nghiên cứu này, nhóm nghiên cứu quan tâm đến sự biến đổi của các đặc trưng theo thời gian. Các đặc trưng có giá trị biến đổi theo thời gian là hành vi nghe nhạc và lịch sử giao dịch. Tuy nhiên các đặc trưng về thông tin khách hàng là dữ liệu tĩnh cũng được thêm vào để bổ sung thông tin cho mô hình dự đoán.

Từ ba loại dữ liệu này, các đặc trưng phù hợp được trích xuất để thiết kế chuỗi thời gian đa biến. Trước khi dựng dữ liệu đầu vào (ma trận D) cho phương pháp được đề xuất, một vectơ đặc trưng chỉ ra hành vi của khách hàng cần được trích xuất cho mỗi tuần. Vectơ đặc trưng này bao gồm tất cả đặc trưng được trích xuất qua quá trình tiền xử lý dữ liệu. Để xây dựng ma trận D, chúng tôi xếp các vectơ đặc trưng này theo chiều dọc như trong công thức dưới đây.

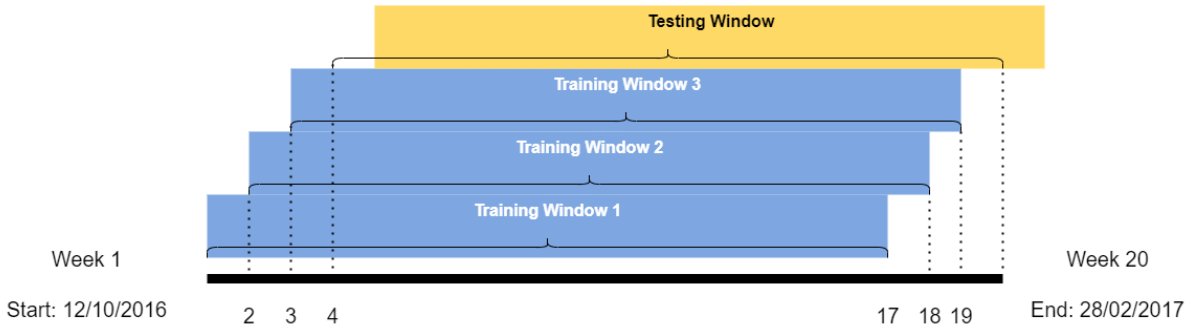
$$D = \begin{pmatrix} \begin{pmatrix} f_1^1(1) & \cdots & f_r^1(1) \\ \vdots & \ddots & \vdots \\ f_1^1(n) & \cdots & f_r^1(n) \end{pmatrix} \\ \vdots \\ \begin{pmatrix} f_1^s(1) & \cdots & f_r^s(1) \\ \vdots & \ddots & \vdots \\ f_1^s(n) & \vdots & f_r^s(n) \end{pmatrix} \end{pmatrix}$$

Tập dữ liệu D bao gồm s chuỗi thời gian đa biến có kích thước r và độ dài n, tức

là mỗi quan sát (khách hàng) ở trong tập dữ liệu bao gồm r chuỗi thời gian có độ dài n, tương ứng với n bước thời gian. Ma trận này được trình bày với hình dạng (s_customers, r_features, n_steps).

3.1.3. Lấy mẫu dữ liệu

Dữ liệu dùng cho việc huấn luyện và kiểm thử hiệu quả mô hình được lấy mẫu trên khoảng thời gian từ ngày 12/10/2016 đến 28/2/2017, tương ứng với dữ liệu chuỗi thời gian 20 tuần. Dữ liệu dùng cho xây dựng mô hình dựa trên lựa chọn các cửa sổ thời gian dài 16 tuần, tương ứng với chuỗi thời gian có độ dài 16 bước thời gian. Các cửa sổ thời gian được đặt tên gồm "Training Window 1", "Training Window 2", "Training Window 3" và "Testing Window", được trình bày ở Hình 3-2.



Hình 3-2. Train - Test Split

Training Window 1 gồm dữ liệu chuỗi thời gian từ tuần 1 đến tuần 17 của khung dữ liệu thực nghiệm, tương tự đối với "Training Window 2" và "Training Window 3". Testing Window kéo dài từ tuần thứ 4 đến tuần thứ 20.

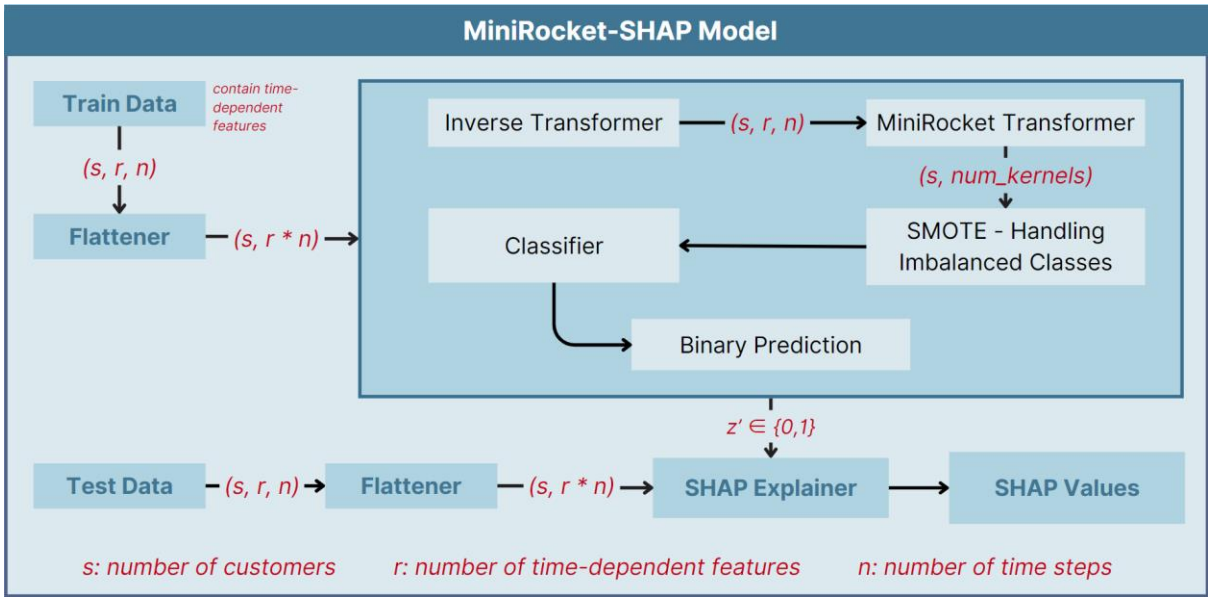
Trong mỗi cửa sổ, dữ liệu thực nghiệm được chọn bằng cách lấy mẫu các khách hàng có gói đăng ký hết hạn trong tuần cuối cùng của cửa sổ. Nếu sau 4 tuần kể từ khi gói đăng ký hết hạn ở tuần cuối cùng của cửa sổ nhưng khách hàng chưa gia hạn (chưa có giao dịch mới), khách hàng được dán nhãn là rời bỏ. Tập dữ liệu huấn luyện là tổng hợp các chuỗi thời gian đa biến (tương đương với các khách hàng) chọn được từ "Training Window 1", "Training Window 2", "Training Window 3". Tập dữ liệu kiểm thử (test) là các chuỗi thời gian đa biến chọn được từ "Testing Window". Sau quá trình lấy mẫu này, thu được bộ dữ liệu thực nghiệm gồm 249086 mẫu ở tập dữ liệu huấn

luyện và 130901 mẫu ở tập dữ liệu kiểm thử.

3.1.4. Mô hình MiniRocket-SHAP

SHAP hoạt động dựa trên lý thuyết trò chơi tập hợp, trong đó các giá trị đặc trưng của một ví dụ dữ liệu đóng vai trò như người chơi trong một liên minh. Các giá trị đặc trưng này thường được đại diện dưới dạng một vectơ liên minh, trong đó mỗi phần tử chỉ ra liệu một đặc trưng có hiện diện hay vắng mặt. Phương pháp SHAP được đề xuất bởi Lundberg and Lee (2017) đã hỗ trợ dữ liệu dạng bảng, hình ảnh và văn bản, tuy nhiên chưa hỗ trợ giải thích dữ liệu chuỗi thời gian.

Do đó, nhóm nghiên cứu đề xuất mô hình MiniRocket-SHAP giúp kết hợp mô hình phân lớp chuỗi thời gian dựa trên MiniRocket và phương pháp SHAP nhằm giải thích kết quả dự đoán. Trong mô hình này, dữ liệu chuỗi thời gian được biến đổi để mỗi bước thời gian được xem như một đặc trưng, sau đó, ở không gian đặc trưng mới này, phương pháp giải thích cục bộ được thực hiện để gán giá trị Shapley cho từng đặc trưng nhằm giải thích đóng góp của nó.



Hình 3-3. MiniRocket-SHAP Model

Việc kết hợp SHAP với một mô hình phân lớp chuỗi thời gian đòi hỏi sự biến đổi trong cấu trúc dữ liệu đầu vào, được mô tả như Hình 3-3. Trong hướng tiếp cận này, ma trận dữ liệu chuỗi thời gian 3 chiều (s, r, n) được làm phẳng thành ma trận 2 chiều (s, $r \times n$). Trong ma trận 2 chiều mới, mỗi bước thời gian $f_r^s(n)$ trong ma trận 3 chiều

được xem như một đặc trưng được giải thích bởi SHAP. Phép biến đổi này được thực hiện bởi "Flattener", áp dụng cho cả tập dữ liệu huấn luyện và tập dữ liệu kiểm thử. Ma trận dữ liệu 2 chiều mới giúp cho KernelSHAP có thể giả lập mô hình hồi quy tuyến tính để tính toán giá trị Shapley cho từng đặc trưng. Các bước phân lớp chuỗi thời gian đa biến được đóng gói vào một "đường ống" (pipeline), đóng vai trò như mô hình được SHAP giải thích. Quá trình phân lớp chuỗi thời gian trong đường ống được miêu tả theo các bước như sau:

Bước 1: Inverse Transformer thực hiện biến đổi ma trận 2 chiều trở lại thành ma trận chuỗi thời gian đa biến 3 chiều (s, r, n).

Bước 2: MiniRocket Transformer thực hiện biến đổi dữ liệu chuỗi thời gian dựa trên kernel tích chập ngẫu nhiên tối thiểu. Sau khi biến đổi thu được dữ liệu với hình dạng (s, num_kernels), trong đó num_kernels là số lượng các kernel tích chập được sử dụng. Dữ liệu này có thể được sử dụng với các mô hình cơ bản như Cây quyết định.

Bước 3: SMOTE được sử dụng để giải quyết vấn đề mất cân bằng lớp. SMOTE là một kỹ thuật liên quan đến việc lấy mẫu quá mức (over-sampling) lớp thiểu số bằng cách tạo các mẫu mới tổng hợp từ lớp thiểu số, nó giải quyết các tập dữ liệu mất cân bằng trong đó lớp thiểu số không được đại diện đầy đủ, giúp nâng cao hiệu suất của mô hình phân lớp (Chawla và cộng sự, 2002).

Bước 4: Huấn luyện mô hình phân loại sử dụng các mô hình cơ bản gồm Cây quyết định, XGBoost, Hồi quy Logistic. Sau quá trình thử nghiệm, kết quả cho thấy các mô hình có độ chính xác gần như tương đương, tuy nhiên Cây quyết định có tốc độ thực thi nhanh nhất, do đó được chọn làm mô hình phân loại chính.

Bước 5: KernelSHAP sử dụng đường ống này như một mô hình dự đoán để thực hiện dự đoán trên tập dữ liệu kiểm thử và có được kết quả phân loại nhị phân.

Sau khi KernelSHAP thực hiện giải thích cục bộ (giải thích kết quả dự đoán lần lượt của từng mẫu) bằng cách mô phỏng lại mô hình hồi quy tuyến tính trên 500 mẫu ở tập dữ liệu kiểm thử, nhóm nghiên cứu thu được giá trị Shapley đại diện cho đóng góp của từng điểm giá trị của các mẫu được sử dụng. Các giá trị Shapley cục bộ thu được có thể được sử dụng để giải thích toàn cục bằng cách tính tổng của các giá trị này. Cụ

thể, việc giải thích toàn cục được thực hiện như sau:

- Độ quan trọng của đặc trưng: trong từng giải thích cục bộ, mỗi giá trị $f_r^s(n)$ của các đặc trưng của mẫu đó được gán một giá trị Shapley. Giá trị Shapley toàn cục của đặc trưng chính là tổng của các giá trị Shapley của đặc trưng tương ứng ở tất cả các mẫu. Trên cơ sở toán học, ta gọi ϕ_r là giá trị Shapley toàn cục của đặc trưng r, $f_r^s(n)$ là đặc trưng được trình bày ở không gian ma trận 2 chiều, $\phi_r^s(n)$ là giá trị Shapley của đặc trưng $f_r^s(n)$, ta có được giá trị Shapley toàn cục được biểu diễn bằng công thức:

$$\phi_r = \sum_{s=1}^{S} \sum_{n=1}^{N} \phi_r^s(n)$$

trong đó N là độ dài chuỗi thời gian tối đa, S là số lượng mẫu được sử dụng để giải thích.

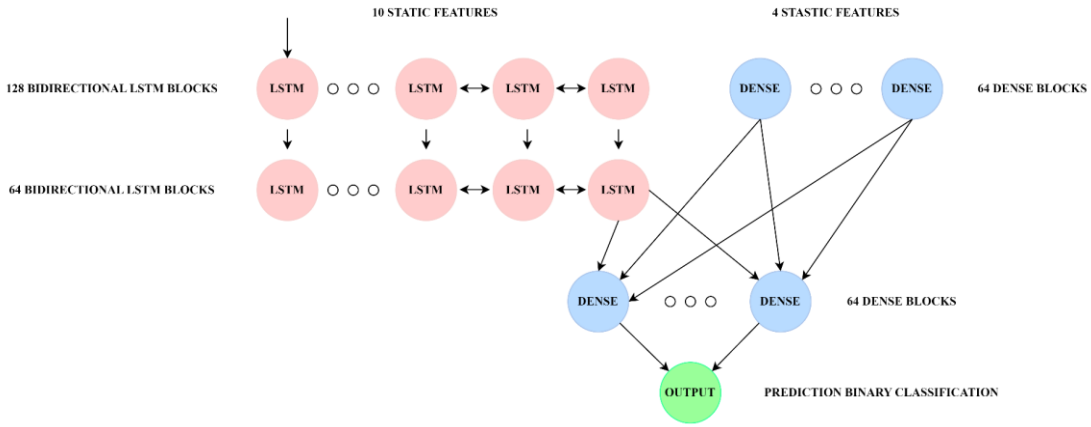
- Độ quan trọng của bước thời gian: ở không gian đặc trưng giả lập nhằm giải thích mô hình, mỗi giá trị $f_r^s(n)$ với n là thứ tự của bước thời gian được xem như một đặc trưng. Nhờ đó ta có được độ quan trọng bước thời gian là tổng của các giá trị Shapley của điểm thời gian tương ứng ở tất cả các mẫu. Gọi ϕ_n là giá trị Shapley toàn cục của bước thời gian n, ta có giá trị Shapley toàn cục được biểu diễn bằng công thức sau:

$$\phi_n = \sum_{s=1}^{S} \sum_{r=1}^{R} \phi_r^s(n)$$

trong đó S là số lượng mẫu tối đa được sử dụng, R là số lượng đặc trưng.

Việc giải thích được độ quan trọng đặc trưng và độ quan trọng của bước thời gian là điểm mới và đột phá của mô hình MiniRocket-SHAP. Độ quan trọng đặc trưng giúp giải thích được các yếu tố ảnh hưởng đến quyết định rời bỏ của khách hàng, từ đó giúp đưa ra các chiến lược giữ chân tương ứng với nguyên nhân rời bỏ. Hơn thế nữa, độ quan trọng của bước thời gian giúp giải thích được thời điểm nào đã ảnh hưởng nhiều nhất đến quyết định rời bỏ của khách hàng.

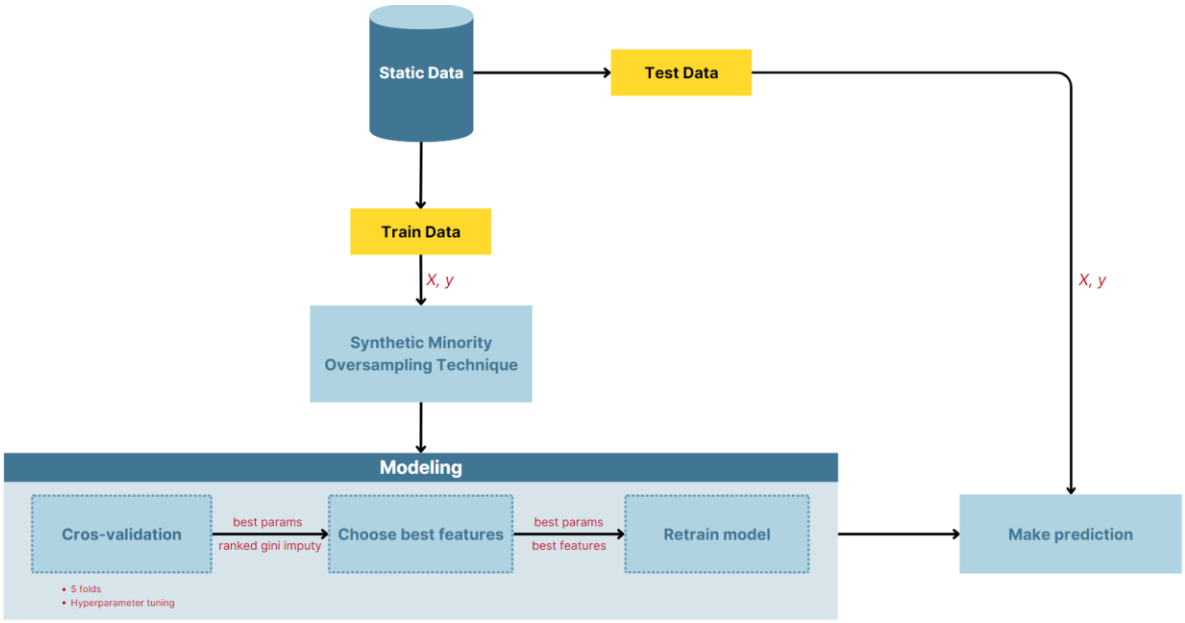
3.1.5. Bi-directional LSTM kết hợp Single Layer Perceptron



Hình 3-4. Mô hình Bidirectional LSTM kết hợp Single Layer Perception

Mô hình Bidirectional LSTM kết hợp Single Layer Perceptron được đề xuất bởi Hyland và cộng sự (2020). Cơ sở cho cách tiếp cận này là Bidirectional LSTM sẽ gồm hai mạng LSTM gọi là forward LSTM và backward LSTM. Mỗi LSTM có trạng thái ẩn và cổng (gate) để điều chỉnh thông tin đầu vào và trạng thái ẩn. LSTM hai chiều sẽ học mối quan hệ giữa các tính năng chuỗi thời gian và lớp đơn perceptron sẽ tập trung vào các tính năng tĩnh của mô hình. SLP là một trong những mạng lưới thần kinh lâu đời nhất và được giới thiệu đầu tiên, được đề xuất bởi Rosenblatt (1958). Perceptron còn được gọi là mạng lưới thần kinh nhân tạo với một lớp đầu vào và một lớp đầu ra, mà không có lớp ẩn. Kiến trúc mô hình bao gồm 1 lớp SLP và 2 lớp LSTM, theo sau là một lớp lớp nối để kết hợp đầu ra từ các lớp LSTM và SLP. Đầu ra kết hợp này sau đó được chuyển sang một lớp dense khác, theo sau là một lớp đầu ra có kích hoạt sigmoid để dự đoán rời bỏ hay không. Các kiến trúc mô hình có thể xem được trong Hình 3-4.

3.1.6. Mô hình RF-Static



Hình 3-5. Mô hình RF-Static

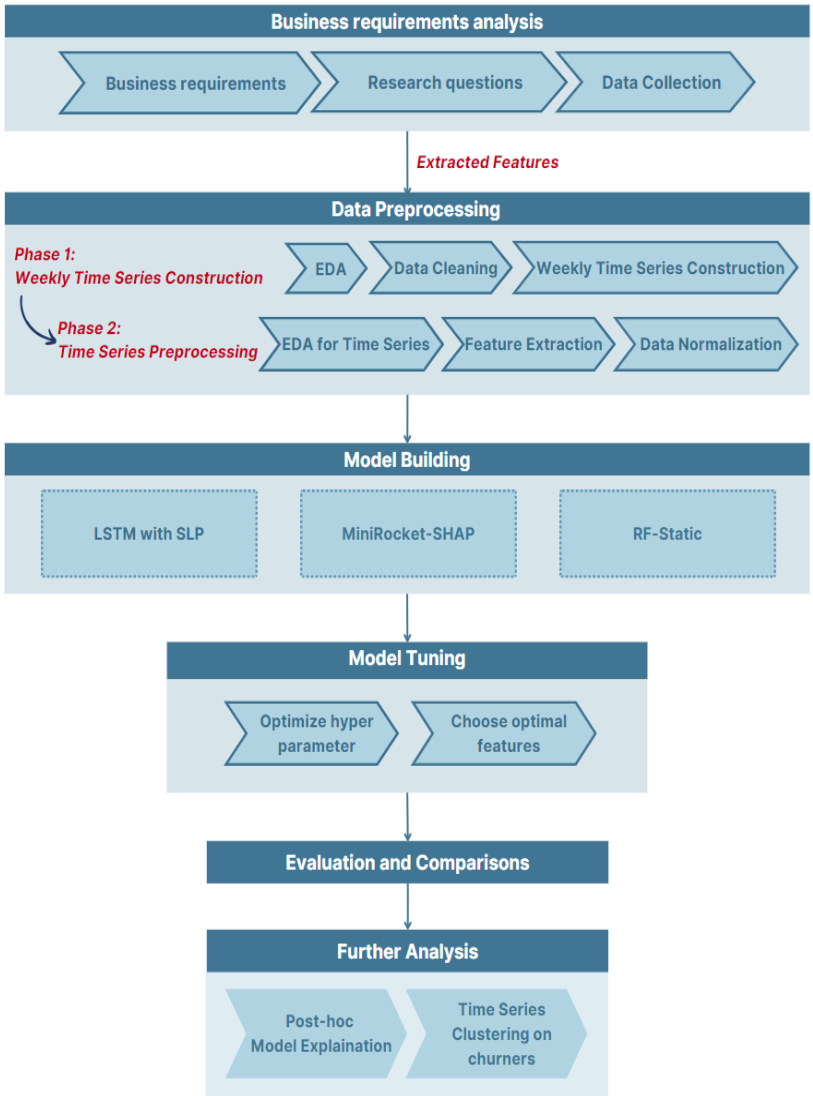
Mô hình RF-Static thực hiện dự đoán với dữ liệu tĩnh thông thường. Dữ liệu tĩnh là tổng giá trị của tất cả các tuần trong mỗi chuỗi thời gian đa biến, riêng đối với các đặc trưng nhân khẩu học được tính bằng trung vị. Dữ liệu tĩnh được chia thành hai phần chính: Dữ liệu huấn luyện và dữ liệu kiểm tra. Dữ liệu huấn luyện được sử dụng để xây dựng mô hình trong quá trình huấn luyện, trong khi dữ liệu kiểm tra được sử dụng để đánh giá hiệu suất của mô hình. Mô hình phân chia dữ liệu thành các đặc trưng (X) và biến mục tiêu (y). Mô hình này sử dụng kỹ thuật SMOTE để giải quyết vấn đề mất cân bằng lớp cho dữ liệu huấn luyện và phương pháp kiểm định chéo (cross validation) để cung cấp tham số tối ưu, cuối cùng là tinh chỉnh mô hình với điểm Gini Impurity để có mô hình tốt nhất. Sau đó mô hình được huấn luyện lại với tham số và nhóm đặc trưng tối ưu được lựa chọn dựa trên xếp hạng điểm Gini Impurity từ mô hình Cây quyết định.

3.2. Thiết kế quy trình thực nghiệm

Mô hình đề xuất bao gồm 4 giai đoạn chính:

Phân tích yêu cầu kinh doanh: Nhóm nghiên cứu xác định các yêu cầu kinh doanh và đặt ra những câu hỏi xoay quanh mục đích và phương pháp churn modelling. Từ những yêu cầu cùng các câu hỏi phản biện được đặt ra, nhóm xác định được các điểm yếu của phương pháp dự đoán rời bỏ truyền thống. Việc khắc phục các điểm yếu

này là cơ sở để thực hiện bài nghiên cứu này với đề xuất phương pháp dự đoán rời bỏ với phân lớp chuỗi thời gian. Sau đó, nhóm tiến hành thu thập và lựa chọn dữ liệu cần thiết cho quá trình nghiên cứu. Ngoài ra, giá trị và chi phí của các hành động như can thiệp rời bỏ hay xác định sai khách hàng rời bỏ cần được xác định để tối ưu mô hình theo yêu cầu kinh doanh.



Hình 3-6. Mô hình đề xuất

Tiền xử lý dữ liệu: Thực hiện phân tích dữ liệu khám phá và tiền xử lý dữ liệu thô từ dữ liệu đã được thu thập trong giai đoạn 1. Giai đoạn này được chia thành 2 giai đoạn nhỏ. Đầu tiên, nhóm xác định và xử lý các giá trị rỗng và ngoại lai trong tập dữ liệu gốc, sau đó tiến hành làm sạch dữ liệu và chuyển đổi dữ liệu từ dạng tĩnh thành dữ liệu chuỗi thời gian theo tuần, mỗi bước thời gian cách nhau 1 tuần. Ở quy trình tiếp theo, sau khi có dữ liệu chuỗi thời gian theo tuần, nhóm tiến hành phân tích khám phá

để hiểu tính chất của dữ liệu và củng cố hiểu biết về vấn đề kinh doanh cần giải quyết. Feature engineering được thực hiện để tăng độ chính xác của mô hình máy học và tạo ra các đặc trưng mới thông qua bước feature extraction. Cuối cùng, bước chuẩn hóa dữ liệu được thực hiện để đảm bảo các đặc trưng có cùng khoảng giá trị.

Huấn luyện và đánh giá: Mô hình MiniRocket-SHAP được xây dựng với nhiều nhóm đặc trưng khác nhau nhằm mục đích so sánh và đánh giá hiệu quả của phương pháp chọn đặc trưng dựa trên giá trị Shapley. Sau đó so sánh mô hình MiniRocket-SHAP với các mô hình LSTM kết hợp Single Layer Perceptron và RF-Static. Các mô hình được đánh giá bằng tập dữ liệu kiểm thử thông qua các chỉ số Precision, Recall, F1 và Log Loss.

Nhận xét và đề xuất sử dụng: Sau khi kết quả phân lớp chuỗi thời gian được đánh giá, nhóm nghiên cứu thực hiện phân tích kết quả phân lớp và xác định yếu tố ảnh hưởng đến hành vi rời bỏ, thời điểm ảnh hưởng đến quyết định rời bỏ. Ngoài ra, kỹ thuật phân cụm dữ liệu chuỗi thời gian cũng được sử dụng để phân cụm các khách hàng được xác định là rời bỏ nhằm đưa ra hiểu biết và đề ra chiến lược phù hợp.

Tóm tắt chương 3

Ở chương 3, nhóm tiến hành xây dựng mô hình dựa trên lý thuyết từ chương 3. Mô hình hoàn chỉnh được xây dựng, chuẩn bị cho quá trình kiểm thử, điều chỉnh và đánh giá. Nhóm đưa ra mô tả chi tiết về kỹ thuật và phương pháp được áp dụng.

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1. Kết quả thực nghiệm

4.1.1. Thu thập và mô tả dữ liệu

Nhóm lựa chọn tập dữ liệu từ hội nghị quốc tế ACM lần thứ 11 về Web Search and Data Mining (Chen và cộng sự, 2018) dữ liệu khách hàng rời bỏ trên nền tảng nghe nhạc đăng ký trực tuyến KKBOX. Dữ liệu bao gồm các bảng:

- Transaction: Ghi nhận các giao dịch của khách hàng. Dữ liệu được sử dụng trong khoảng thời gian từ 01/01/2015 đến ngày 28/02/2017. Với 10 thuộc tính được mô tả như sau:

Bảng 4-1. Mô tả các thuộc tính trong bảng Transaction

STT	Thuộc tính	Mô tả	Kiểu dữ liệu	Các giá trị
1	msno	Mã định danh của đơn đặt hàng khách hàng khách hàng là duy nhất	object	Gồm 974,578 mã duy nhất với từng khách hàng
2	payment_method_id	Phương thức thanh toán của khách hàng được mã hóa	int64	Gồm 26 giá trị được mã hóa thành: 10, 11, 14, 16, 17, 18, 19, 20, 21, 23, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41.
3	payment_plan_days	Thời hạn của gói thành viên được tính theo ngày	int64	

4	plan_list_price	Giá niêm yết theo kế hoạch.	int64	
5	actual_amount_paid	Số tiền trả thực tế.	int64	
6	is_auto_renew	Khách hàng chọn tự động gia hạn hoặc không	int64	1: có tự động gia hạn 0: không tự động gia hạn
7	transaction_date	Ngày xảy ra giao dịch	int64	
8	membership_expire_date	Ngày hết hạn thành viên	int64	
9	is_cancel	Người dùng có hủy tư cách thành viên trong giao dịch này hay không	int64	1: hủy, 0: không hủy

- Userlog: Nhật ký người dùng hàng ngày mô tả hành vi nghe của người dùng. Dữ liệu được thu thập cho đến ngày 28/02/2017. Với 9 thuộc tính được mô tả như sau:

Bảng 4-2. Mô tả các thuộc tính trong bảng Userlog

STT	Thuộc tính	Mô tả	Kiểu dữ liệu	Các giá trị
1	msno	Mã định danh của đơn đặt hàng khách hàng khách hàng là duy nhất	object	Gồm 974,578 mã duy nhất với từng khách hàng
2	date	Ngày ghi nhận hành động của khách hàng	int64	

3	num_25	Số bài hát được phát dưới 25% thời lượng bài hát	int64
4	num_50	Số bài hát được phát từ 25% đến 50% thời lượng bài hát	int64
5	num_75	Số bài hát được phát từ 50% đến 75% thời lượng bài hát	int64
6	num_985	Số bài hát được phát từ 75% đến 98.5% thời lượng bài hát	int64
7	num_100	Số bài hát được phát trên 98,5% thời lượng bài hát	int64
8	num_unq	Số bài hát được nghe. Khác với số lượt nghe bài hát	int64
9	total_secs	Tổng thời gian nghe nhạc tính trên đơn vị giây	int64

- Member: Thông tin người khách hàng của KKBOX. Với 7 thuộc tính được mô tả như sau:

Bảng 4-3. Mô tả thuộc tính trong bảng Member

STT	Thuộc tính	Mô tả	Kiểu dữ liệu	Các giá trị
1	msno	Mã định danh của đơn đặt hàng khách hàng khách hàng là duy nhất	object	Gồm 6,769,473 mã duy nhất với từng khách hàng
2	city	Thành phố khách hàng sinh sống	int64	Gồm 22 giá trị được mã hóa thành các số từ 1 đến 22
3	bd	Số tuổi của khách hàng	int64	

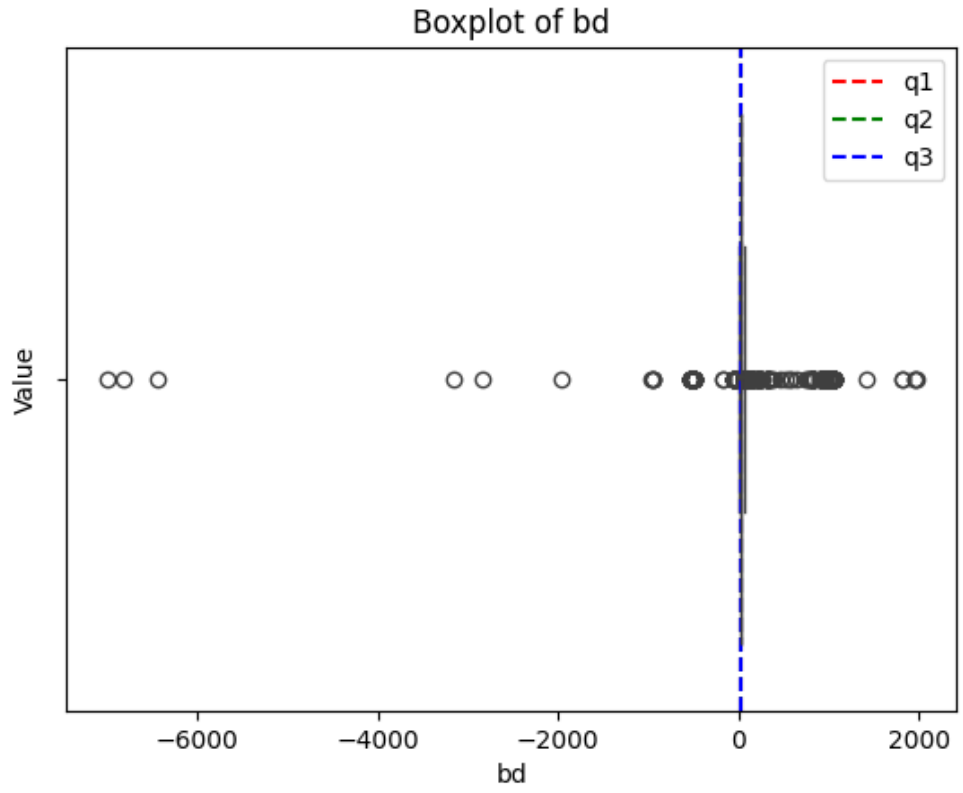
4	gender	Giới tính	object	Male: Nam, Female: Nữ và một số giá trị bỏ trống
5	registered_via	Phương thức đăng ký	int64	Gồm 20 giá trị được mã hóa thành các số -1 và 1 đến 19
6	registration_init_time	Thời điểm bắt đầu đăng ký	int64	
7	expiration_date	Thời gian hết hạn	int64	

4.1.2. Tiền xử lý dữ liệu

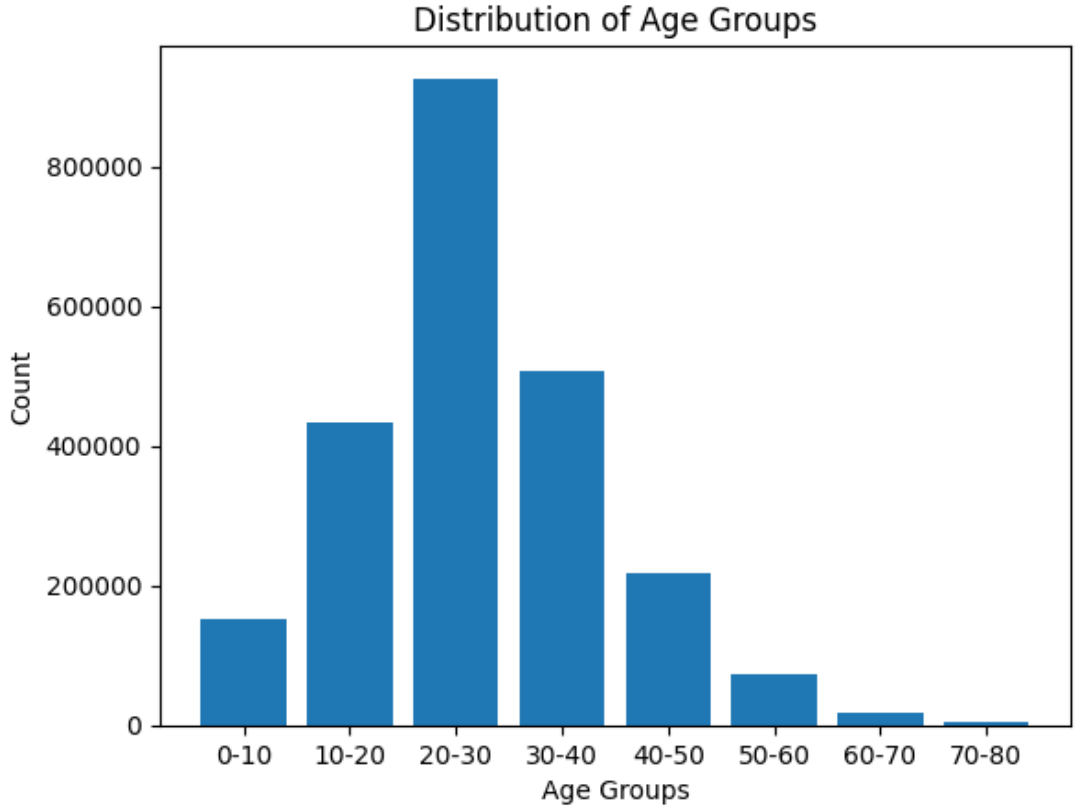
Giai đoạn 1: Phân tích khám phá dữ liệu và làm sạch dữ liệu ban đầu

Tiến hành khám phá dữ liệu cho thấy sự tự không tương đồng về các mã khách hàng xuất hiện ở các bảng, do đó cần lần lượt loại bỏ các khách hàng không tồn tại ở tất cả các bảng. Ngoài ra, quá trình kiểm tra cho thấy dữ liệu tồn tại các giá trị null, các giá trị không phù hợp và có nhiều điểm ngoại lại. Các giá trị null chỉ tồn tại ở thuộc tính "gender" của bảng dữ liệu "Member", tuy nhiên, "gender" được xem như là một yếu tố quan trọng để phân loại khách hàng vậy nên cần loại bỏ các dòng giá trị của các khách hàng không phù hợp và giữ lại đặc trưng này để không ảnh hưởng đến mô hình.

Ngoài ra, đặc trưng bd (tuổi của khách hàng) ở bảng dữ liệu "Member" cũng chứa nhiều giá trị ngoại lai như được trình bày ở Hình 4-1. Đặc trưng này được làm sạch bằng cách loại bỏ các điểm dữ liệu âm và nhóm tuổi khách hàng thành các nhóm tuổi nhằm giảm ảnh hưởng của giá trị ngoại lai, đổi tên thành "age_group". Tuổi của khách hàng được phân vào các nhóm gồm: "người trẻ" (10-20), "người lao động" (20-30), "trung niên"(30-40), "thành niên" (40-50), "lão niên" (50-60) và nhóm "khác" (thuộc các tuổi còn lại). Các nhóm tuổi được mã hóa theo thứ tự từ 0 - 5. Ngoài ra, các mẫu dữ liệu có tuổi thuộc 0 - 10 tuổi đã bị loại bỏ do không phù hợp với giả định khách hàng có thể đăng ký tài khoản và tự chi trả.

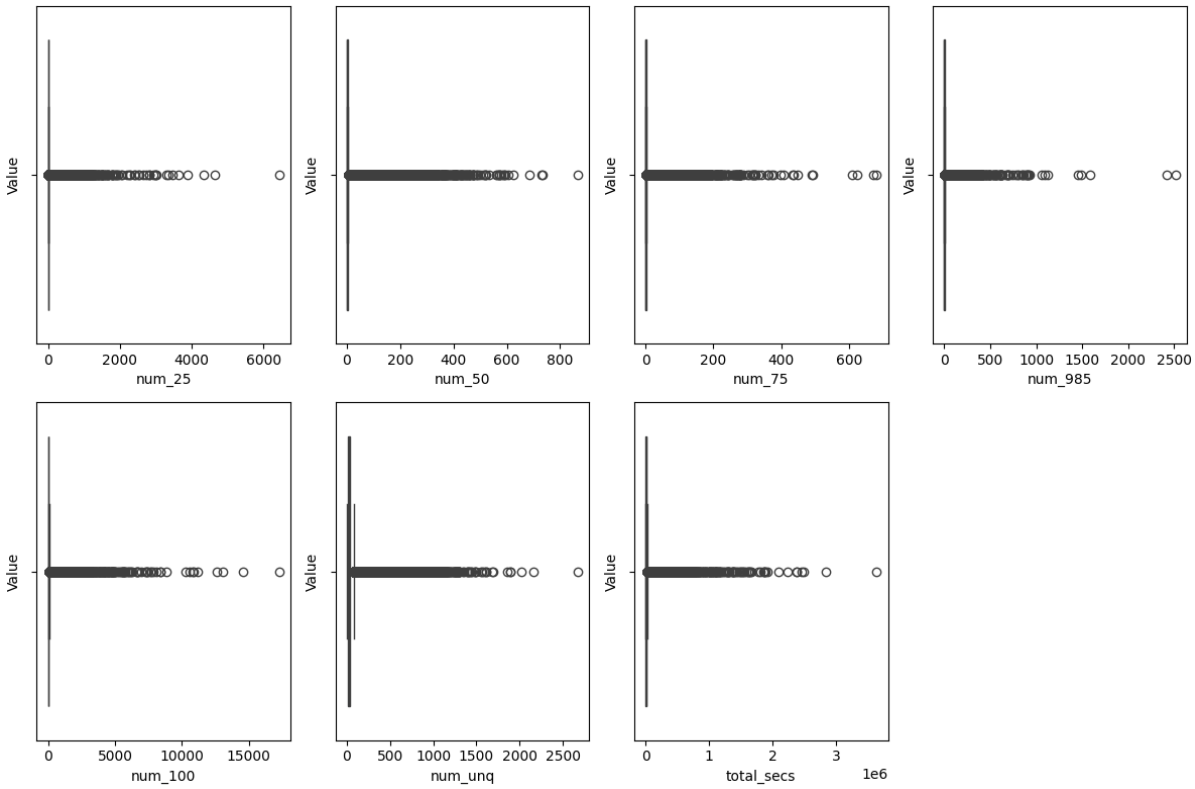


Hình 4-1. Biểu đồ hộp phân bố tuổi thông qua biến bd



Hình 4-2. Biểu đồ cột mô tả phân phối nhóm tuổi

Đối với bảng dữ liệu về nhật ký hành vi khách hàng (Userlog), các biến về hành vi khách hàng đều chứa nhiều giá trị ngoại lai như được trình bày bằng biểu đồ hộp ở . Dựa vào hiểu biết về các nền tảng nghe nhạc trực tuyến, nhóm nghiên cứu xác định các giá trị ngoại lai là các giá âm hoặc các giá trị lớn bất thường.



Hình 4-3. Biểu đồ các giá trị

Để loại bỏ các điểm ngoại lai khỏi tập dữ liệu, phương pháp xử lý sử dụng giá trị IQR được áp dụng chung cho tất cả các đặc trưng có chứa giá trị ngoại lai. Phương pháp này xác định giới hạn dưới và giới hạn trên, sau đó giá trị ngoại lai được xác định là các giá trị nằm ngoài phạm vi này, phạm vi này được xác định là từ $Q1 - 1.5 * IQR$ đến $Q3 + 1.5 * IQR$. Dựa trên hiểu biết về hành vi nghe nhạc, tất cả các giá trị lớn bất thường được quy về giá trị $Q3 + 1.5 * IQR$ để giảm thiểu số điểm ngoại lai nhưng không làm ảnh hưởng ý nghĩa của dữ liệu.

Xây dựng biến mới

Khách hàng được xác định là rời bỏ trong điều kiện chưa gian hạn sau 4 tuần tính từ giá trị "membership_expire_date" của transaction cuối cùng trong khung thời gian thực nghiệm. Ở bảng dữ liệu "Transaction", dữ liệu ban đầu chưa có các biến mang

ý nghĩa theo tuần vậy nên cần xây dựng biến mới là "week", đóng vai trò như chỉ mục (index) thứ tự cho dữ liệu chuỗi thời gian, và các biến khác với giá trị có thể biến đổi mỗi tuần. Mỗi ghi nhận về hành vi khách hàng hoặc giao dịch khách hàng sẽ được đánh số từ 1 đến 16 tương đương với tuần diễn ra giao dịch hoặc hành vi. Thông qua các đặc trưng có thể biến đổi theo tuần này, dữ liệu chuỗi thời gian đa biến được tạo ra như được mô tả ở ma trận D (được đề cập ở phần 3.1.2). Sau khi tạo ra các biến mới có thể biến đổi theo tuần ở bảng dữ liệu "Transaction", các bảng dữ liệu được gộp lại thành một tập dũ liệu thực nghiệm chung, với các biến được trình bày ở Bảng 4-4. Lưu ý rằng dù các đặc trưng "city", "age_group", "gender" và "registered_via" tuy hiếm khi hoặc không biến đổi theo tuần, nhưng có thể kết hợp với dũ liệu chuỗi thời gian trong mô hình phân lớp để đạt được kết quả tốt hơn.

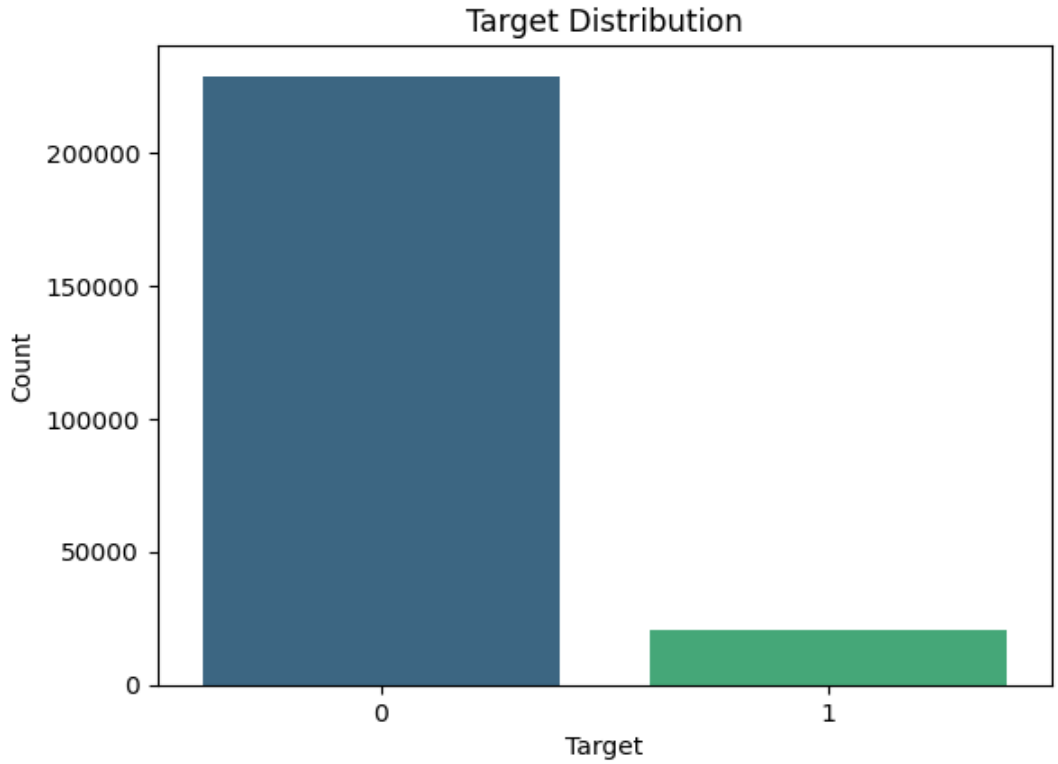
Bảng 4-4. Đặc trưng và ý nghĩa của các biến

Đặc trưng	**Ý nghĩa**
num_25	Số bài hát nghe dưới 25% thời lượng
num_50	Số bài hát nghe dưới 50% thời lượng
num_75	Số bài hát nghe dưới 75% thời lượng
num_985	Số bài hát nghe dưới 98.5% thời lượng
num_100	Số bài hát nghe dưới 100% thời lượng
num_unq	Số lượng bài hát riêng biệt đã nghe
total_secs	Tổng thời gian nghe nhạc
actual_amount_paid	Số tiền thực tế đã trả cho mỗi giao dịch
diff_actual_plan_paid	Số tiền thực tế đã trả cho mỗi giao dịch trừ đi số tiền cần trả trên lý thuyết
make_cancellation	Hủy gói đăng ký
auto_renew	Tự động gia hạn
city	Thành phố
age_group	Nhóm tuổi
gender	Giới tính

registered_via	Phương thức đăng ký tài khoản
weeks_since_registration	Số tuần đếm từ ngày đăng ký đến thời điểm của bước thời gian

Giai đoạn 2: Tiền xử lý dữ liệu chuỗi thời gian

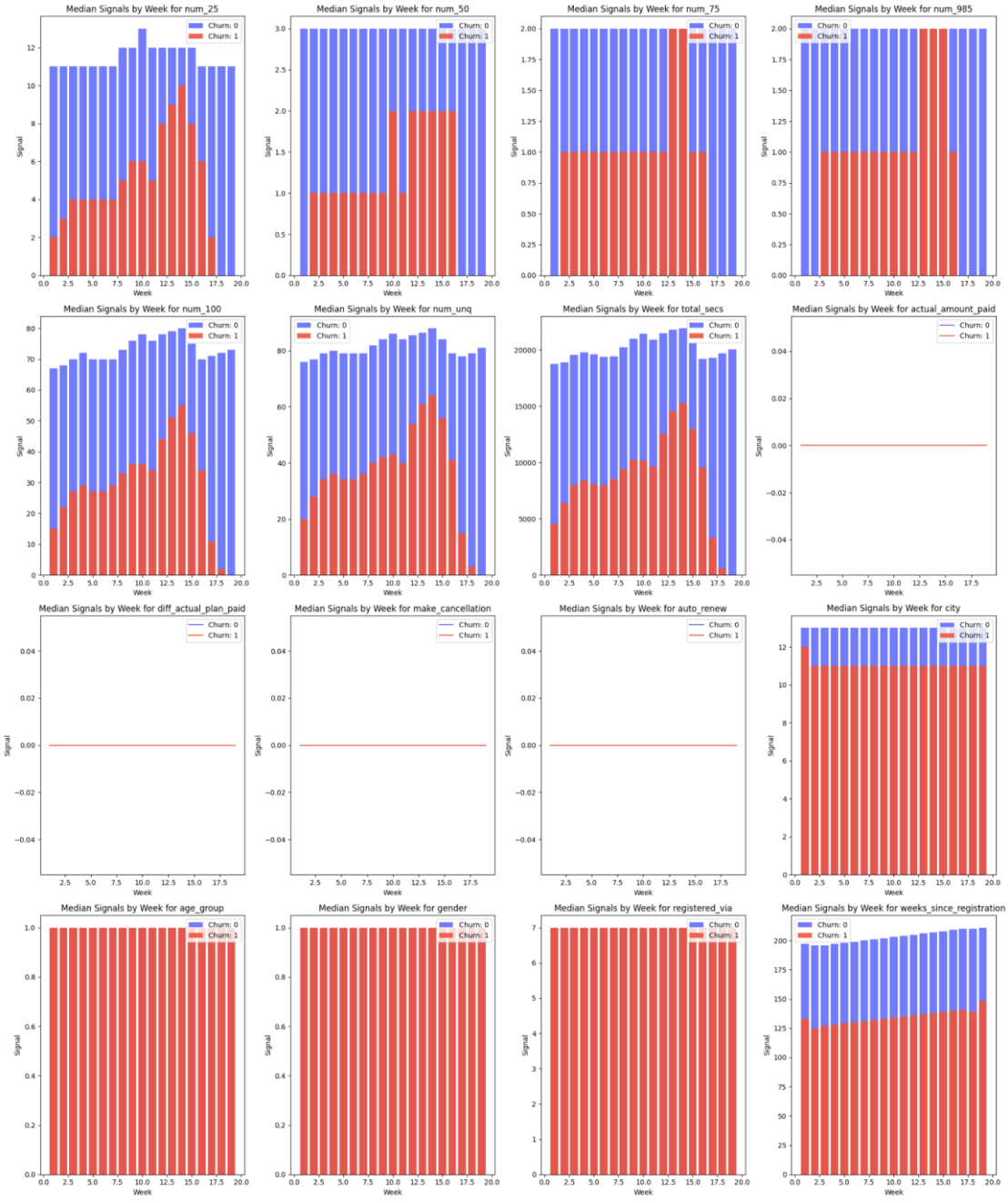
Giai đoạn tiếp theo của quá trình tiền xử lý dữ liệu được thực hiện trên giả định việc tạo ra dữ liệu chuỗi thời gian đa biến đã thêm các điểm dữ liệu mới, khiến cho phân phối của dữ liệu đã thay đổi so với ban đầu. Do đó, việc phân tích khám phá dữ liệu được thực hiện lại cho dữ liệu chuỗi thời gian đa biến mới với hướng tiếp cận khác so với dữ liệu dạng bảng. Sau khi có được r chuỗi thời gian đa biến, việc chia tập dữ liệu huấn luyện và kiểm thử được thực hiện như đã miêu tả ở hình 3-2. Sau quá trình chọn mẫu, thu được 249086 chuỗi thời gian đa biến ở tập dữ liệu huấn luyện (train) và 130901 chuỗi thời gian đa biến ở dữ liệu kiểm thử (test).



Hình 4-4. Biểu đồ phân phối của lớp Churn và Non-Churn

Hình 4-4 cho thấy phân phối không đồng đều của 2 lớp "Churn" và "Non-churn" ở tập dữ liệu huấn luyện. Số lượng khách hàng được dán nhãn là "Churn" chỉ chiếm 9%. Tương tự, ở tập dữ liệu kiểm thử, số lượng khách hàng được dán nhãn là "Churn"

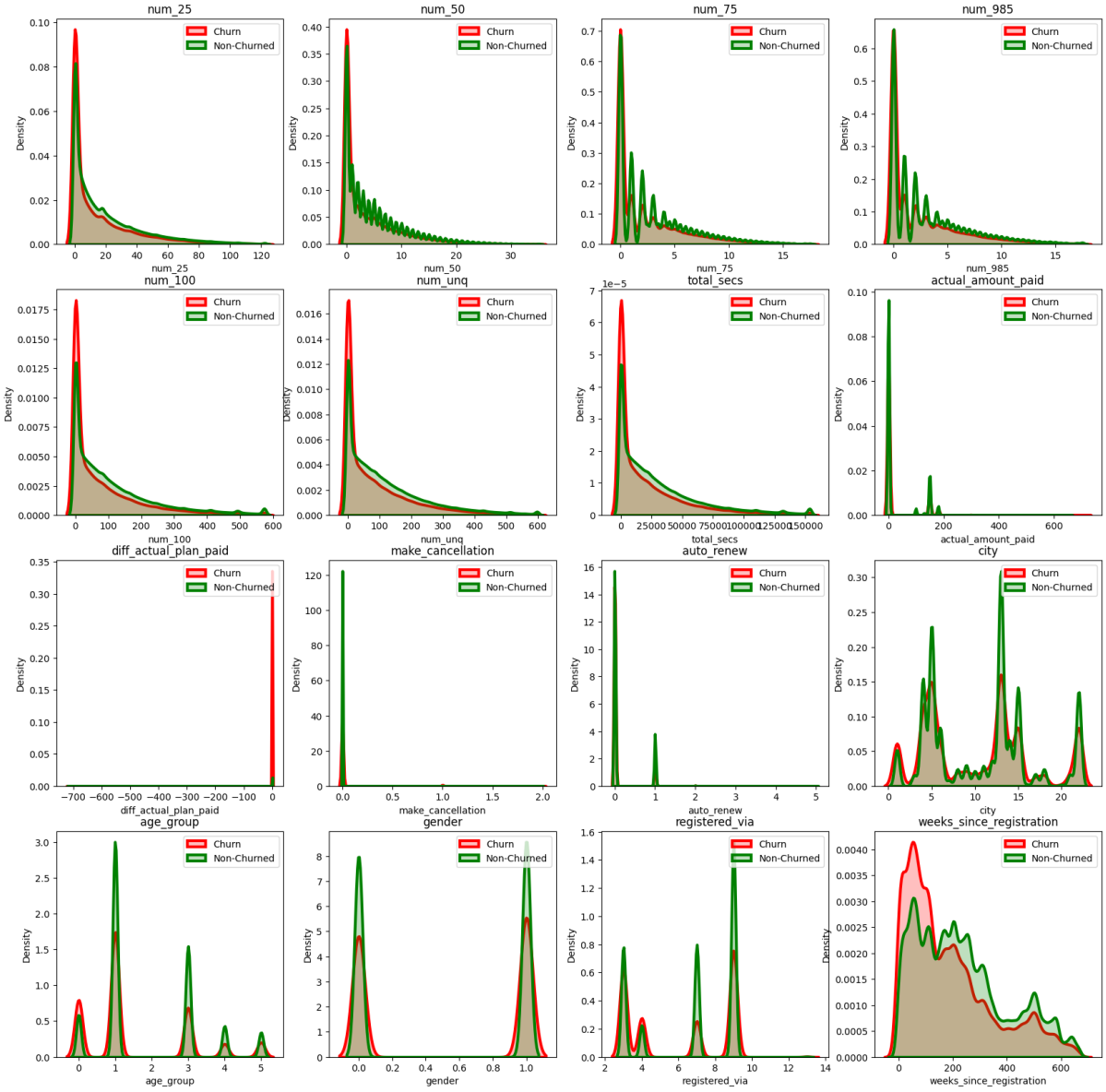
chiếm 6%. Điều này đặt ra bài toán phân lớp trên dữ liệu mất cân bằng lớp. Phương pháp nhằm giải quyết vấn đề này dựa trên kỹ thuật SMOTE được trình bày ở phần sau.



Hình 4-5. Trực quan hóa dữ liệu hai lớp Churn và Non-Churn

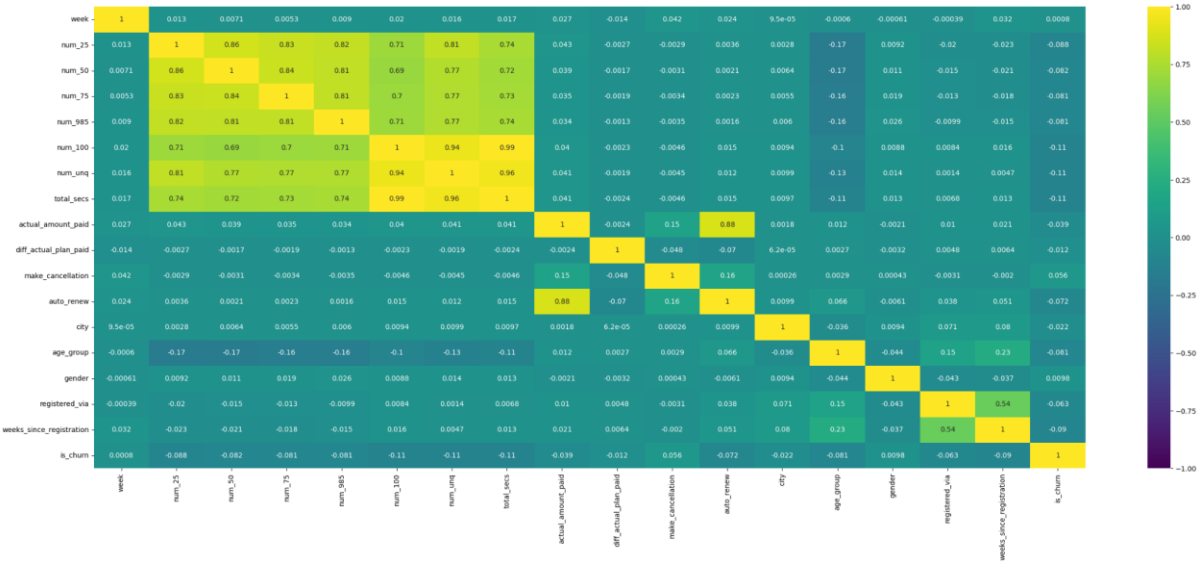
Trực quan hóa trung vị của tín hiệu giữa hai lớp "Churn" và "Non-churn" cho thấy có sự khác biệt giữa hai lớp trong các đặc trưng về hành vi như num_25, num_50, num_75, num_985, num_100, num_unq, total_secs. Sự khác biệt này cho thấy hành vi nghe của nhóm "Churn" có khác biệt với nhóm "Non-churn", tuy nhiên sự biến động của trung vị theo thời gian có hình dáng gần giống nhau. Các đặc trưng khác không cho

thấy sự khác biệt trong tín hiệu trung vị giữa hai lớp là do các đặc trưng này có giá trị nhị phân hoặc có phương sai rất nhỏ, khác tính chất với các đặc trưng về hành vi.



Hình 4-6. Phân phối dữ liệu của 2 lớp Churn và Non-churn

Mặt khác, khi dựa vào phân phối tổng thể giữa hai lớp "Churn" và "Non-churn" của các đặc trưng, như được trình bày ở hình 4-5, có thể thấy phân phối giữa hai lớp ở các đặc trưng về hành vi "num_25", "num_50", "num_75", "num_985", "num_unq", "total_secs" khá tương đồng với nhau. Điều này cho thấy khả năng hành vi giữa hai lớp có ít sự khác biệt (1). Nhóm nghiên cứu đề xuất sử dụng phương pháp t-test để kiểm tra liệu có sự khác biệt đáng kể về mặt thống kê giữa giá trị trung bình của hai nhóm và có mối liên hệ giữa chúng hay không.



Hình 4-7. Tương quan Spearman giữa các đặc trưng

Ngoài ra, giữa các đặc trưng về hành vi có sự tương quan cao dựa trên phương pháp Hệ số tương quan xếp hạng của Spearman (2), được trình bày trên hình 4-6. Dựa trên luận điểm (1) và (2), nhóm nghiên cứu lập luận rằng các đặc trưng về hành vi khách hàng có thể tạo ra nhiễu cho mô hình phân loại, do đó nhóm nghiên cứu giới hạn phạm vi xử lý đặc trưng (feature engineering) trên các đặc trưng này.

Trích xuất đặc trưng

Sau khi hiểu dữ liệu và xác định các đặc trưng có thể gây nhiễu, nhóm nghiên cứu thực hiện trích xuất đặc trưng mới từ các đặc trưng này nhằm trích lọc các thông tin mới và giảm sự phụ thuộc giữa các đặc trưng (sự tương quan giữa các đặc trưng). Phương pháp trích xuất đặc trưng mới nhắm tới mục tiêu nhấn mạnh tính chất thời gian của các đặc trưng hay các chuỗi thời gian. Các đặc trưng về thời gian được trích xuất gồm 3 loại như sau:

Đặc trưng về độ trễ:

- Lag_1: Được tạo ra bằng cách dịch chuyển đặc trưng gốc đi một bước thời gian cho mỗi người dùng và điền các giá trị thiếu bằng trung vị.
- Diff: Tính bằng cách tính hiệu giữa đặc trưng gốc và giá trị trễ của nó.

Đặc trưng động:

- roll_mean_3: Trung bình trượt qua một cửa sổ 3 bước thời gian.

- roll_mean_6: Trung bình trượt qua một cửa sổ 6 bước thời gian.
- roll_mean_9: Trung bình trượt qua một cửa sổ 9 bước thời gian.

Sau khi trích xuất, các giá trị bị thiếu trong các đặc trưng trung bình trượt được điền bằng giá trị kế tiếp trong chuỗi.

Thống kê cho dữ liệu chuỗi:

- mean: Tính bằng cách tính trung bình của đặc trưng cho mỗi tuần.
- mean_diff: Tính bằng cách tính trung bình của sự khác biệt giữa các giá trị liên tiếp của đặc trưng cho mỗi tuần.
- med: Tính bằng cách tính trung vị của đặc trưng cho mỗi tuần.
- std: Tính bằng cách tính độ lệch chuẩn của đặc trưng cho mỗi tuần.
- skew: Tính bằng cách tính độ lệch của phân phối đặc trưng cho mỗi tuần.
- kurt: Tính bằng cách tính độ nhọn của phân phối đặc trưng cho mỗi tuần.
- min: Tính bằng cách tính giá trị tối thiểu của đặc trưng cho mỗi tuần.
- max: Tính bằng cách tính giá trị tối đa của đặc trưng cho mỗi tuần.

Sau quá trình trích xuất và xây dựng đặc trưng mới, đồng thời loại bỏ một số đặc trưng mới nhưng có quá nhiều giá trị rỗng, tổng cộng thu được tập dữ liệu có 85 đặc trưng được tạo ra theo phương pháp như trên. Do các đặc trưng không nằm chung một khoảng giá trị, tất cả đặc trưng được chuẩn hóa sử dụng StandardScaler. Các giá trị được biến đổi dựa trên điểm z, công thức của phương pháp chuẩn hóa được trình bày như sau:

$$z = \frac{x - \mu}{\sigma}$$

Trong đó:

- z : là giá trị mới.
- x : là giá trị gốc ban đầu.
- μ : là giá trị trung bình (mean).
- σ : là độ lệch chuẩn.

4.1.3. Huấn luyện và đánh giá mô hình phân lớp

Sau khi thu được tập dữ liệu gồm các đặc trưng nhấn mạnh tính chất thời gian, bước tiếp theo là sử dụng tập dữ liệu huấn luyện và kiểm thử mới thu được để xây dựng các mô hình MiniRocket-SHAP, LSTM kết hợp SLP và mô hình RF-Static. Trong đó, mô hình MiniRocket-SHAP và RF-Static được tối ưu bằng phương pháp GridSearchCV, bao gồm kiểm định chéo (cross-validation) và tìm kiếm siêu tham số tối ưu (hyperparameter tuning).

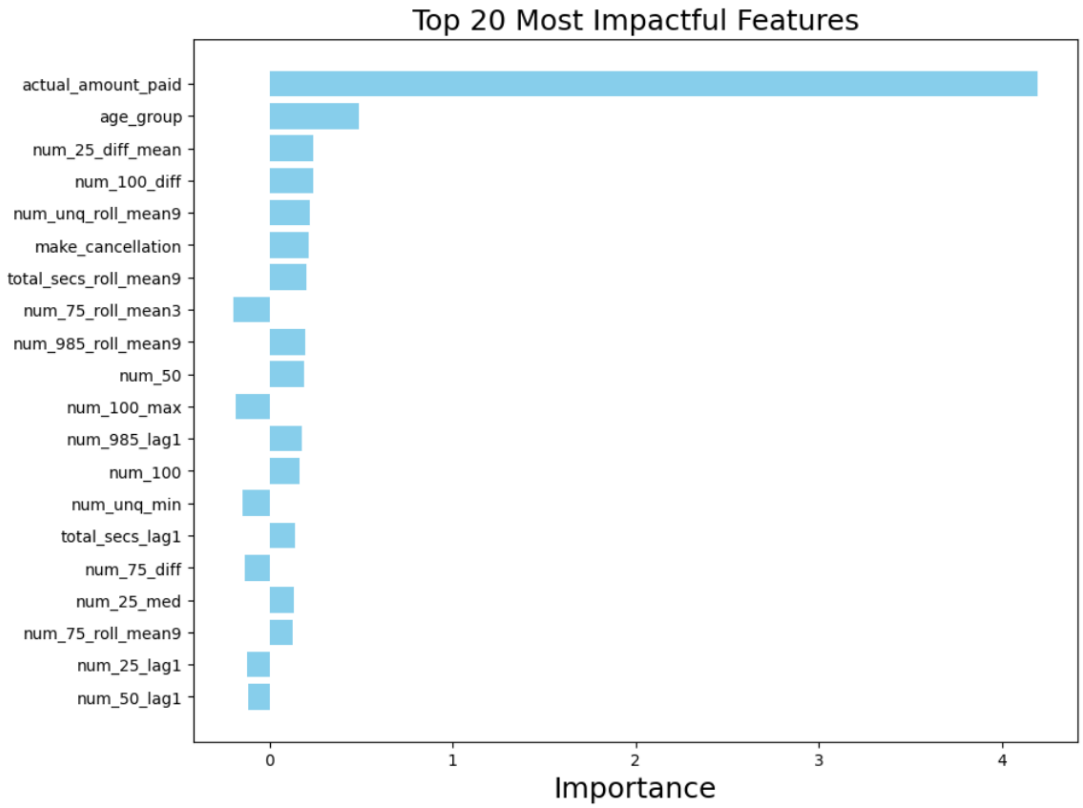
MiniRocket-SHAP

Trong các thành phần cấu thành mô hình MiniRocket-SHAP, siêu tham số của Cây quyết định được tối ưu bằng cách tìm kiếm và đánh giá từng tổ hợp trong không gian siêu tham số, kết hợp với kiểm định chéo. Các tham số trong không gian tìm kiếm bao gồm:

Bảng 4-5. Tham số tìm kiếm và tham số tối ưu

Siêu tham số tìm kiếm	Siêu tham số tối ưu
n_estimators: [100, 250, 500, 1000], max_depth: [10,15,20]	n_estimators: 500, max_depth: 20

Chọn đặc trưng



Hình 4-8. Top 20 đặc trưng có tác động nhất

Mô hình phân lớp chuỗi thời gian MiniRocket-SHAP được thiết kế kết hợp với phương pháp SHAP, giúp giải thích được kết quả phân lớp, như trình bày ở hình 4-8. Từ kết quả độ quan trọng đặc trưng của mô hình đã được tối ưu tham số, nhóm nghiên cứu đặt ra một ngưỡng giá trị quan trọng tối thiểu và chọn các đặc trưng có giá trị quan trọng cao hơn ngưỡng đó. Ngưỡng tối thiểu này giúp chọn ra hai mươi đặc trưng có độ quan trọng cao nhất, được trình bày ở hình 4-8. Do ngưỡng giá trị này có tính ngẫu nhiên nên để đảm bảo chọn ra được nhóm đặc trưng tối ưu, nhóm nghiên cứu tiếp tục sử dụng phương pháp Rank-based Forward Feature Selection để chọn ra nhóm đặc trưng tối ưu từ hai mươi đặc trưng có được. Thuật toán Rank-based Forward Feature Selection được thực hiện như sau:

Bước 1: Khởi tạo không gian mẫu R^{20} gồm 20 đặc trưng được xếp hạng độ quan trọng dựa trên giá trị Shapley.

Bước 2: Bắt đầu với không gian đặc trưng con chứa chỉ một đặc trưng R^1, bắt đầu với đặc trưng có giá trị quan trọng cao nhất là “actual_amount_paid”.

Bước 3: Đánh giá hiệu suất của mô hình sử dụng chỉ đặc trưng hiện tại bằng F1

score.

Bước 4: Thêm đặc trưng có giá trị Shapley lớn thứ 2 vào không gian đặc trưng con R^2 và đánh giá lại hiệu suất của mô hình.

Bước 5: Tiếp tục thêm đặc trưng có giá trị Shapley lớn thứ k vào không gian đặc trưng con R^k và đánh giá lại hiệu suất của mô hình cho đến khi tất cả các đặc trưng đã được thêm vào không gian đặc trưng con và đánh giá.

Bước 6: Chọn ra nhóm đặc trưng có điểm F1 lớn nhất từ các không gian đặc trưng con đã đánh giá.

Sau khi thực hiện thuật toán Rank-based Forward Feature Selection, các đặc trưng trong bảng 4-6 cho ra kết quả phân lớp tốt nhất và được chọn làm nhóm đặc trưng tối ưu sử dụng cho mô hình cuối.

Bảng 4-6. Nhóm đặc trưng được lựa chọn bằng thuật toán Rank-based Forward Feature Selection

Đặc trưng	Giải thích
actual_amount_paid	Số tiền thực tế đã thanh toán
age_group	Nhóm tuổi
num_25_diff_mean	Chênh lệch trung bình của num_25
num_100_diff	Chênh lệch của num_100
num_unq_roll_mean9	Cửa sổ trượt 9 bước của num_unq
make_cancellation	Đã hủy bỏ giao dịch (True/False)
total_secs_roll_mean9	Cửa sổ trượt 9 bước của total_secs
num_75_roll_mean3	Cửa sổ trượt 3 bước của num_75

Sau khi chọn được nhóm đặc trưng tối ưu, việc huấn luyện và đánh giá mô hình phân lớp chuỗi thời gian được thực hiện lại với dữ liệu đầu vào gồm nhóm đặc trưng này, cùng các siêu tham số tối ưu. Ngoài ra, mô hình MiniRocket-SHAP cũng được huấn luyện với các không gian đặc trưng khác với mục đích so sánh và đánh giá hiệu quả của phương pháp chọn đặc trưng từ giá trị Shapley. Các không gian đặc trưng này gồm:

- R^8: không gian đặc trưng tối ưu được chọn.

- D^{16}: không gian đặc trưng gốc.
- E^{85}: không gian đặc trưng gồm tất cả đặc trưng được tạo sau quá trình trích xuất đặc trưng mới.

Trong bài nghiên cứu này, để trình bày tốt hiệu quả các mô hình trong bài toán mất cân bằng lớp, việc chọn đúng chỉ số để đánh giá là rất quan trọng. Chỉ số F1 là thích hợp để đánh giá bài toán mất cân bằng lớp bởi vì có sự kết hợp giữa hai chỉ số Precision và Recall. Do đó, chỉ số F1 được chọn là chỉ số chính để đánh giá tổng quan và so sánh các mô hình, kết hợp xem xét các chỉ số khác như Accuracy, Precision, Recall, Log Loss.

Bảng 4-7. Kết quả mô hình thông qua các chỉ số

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Log Loss**
MiniRocket-SHAP (R^8)	**0.95**	0.82	0.77	**0.79**	1.74
MiniRocket-SHAP (E^{85})	0.87	0.62	0.57	0.58	4.7
MiniRocket-SHAP (D^{16})	0.65	0.66	0.54	0.48	12.52

Các mô hình MiniRocket-SHAP trên được huấn luyện bằng các tham số của từng thành phần trong mô hình được trình bày trong bảng sau:

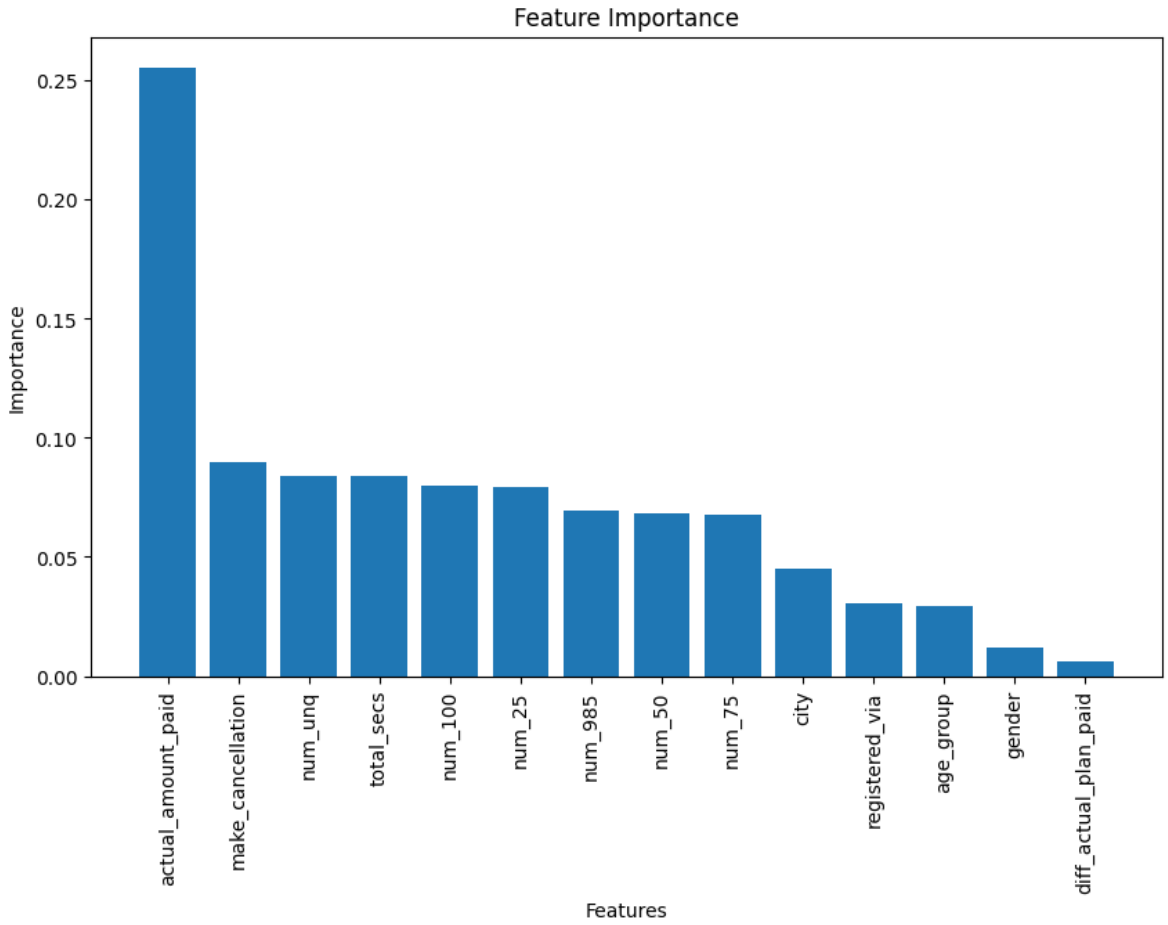
Bảng 4-8. Bảng siêu tham số tối ưu trên mô hình MiniRocket-SHAP

Thành phần	Siêu tham số tối ưu
MiniRocket Transformer	num_kernel: 5000
SMOTE	sampling_strategy: ‘minority'
Cây quyết định Classifier	n_estimators: 500, max_depth: 20

RF-Static

Ngoài ra, nhóm nghiên cứu cũng thực hiện xây dựng mô hình dự đoán rời bỏ truyền thống dựa trên dữ liệu tĩnh nhằm mục đích so sánh. Mỗi đặc trưng trong tập dữ liệu này được tạo ra bằng cách lấy tổng hoặc trung bình, trung vị của tất cả các tuần trong khung thời gian thực nghiệm. Quy trình xây dựng mô hình này được mô tả như

Hình 4-9.



Hình 4-9. Biểu đồ thể hiện độ quan trọng của đặc trưng trong mô hình RF-Static

Sau khi thực hiện cross-validation bằng phương pháp GridSearchCV với 5 "nếp gấp" (fold), nhóm nghiên cứu thu được các tham số tối ưu và độ quan trọng đặc trưng xếp hạng dựa trên chỉ số Gini Impurity, như Hình 4-9. Sau đó, có chín đặc trưng quan trọng được lựa chọn để huấn luyện lại mô hình dựa trên các tham số tối ưu. Các đặc trưng này gồm: "actual_amount_paid", "make_cancellation", "num_unq", "total_secs", "num_100", "num_25", "num_50", "num_985" và "num_75".

LSTM kết hợp SLP

Bên cạnh đó, mô hình LSTM kết hợp Single Layer Perceptron cũng được sử dụng với mục đích so sánh và đánh giá. Thiết kế của mô hình học sâu này được trình bày ở hình 3-4. Trong đó, mạng LSTM hai chiều (Bi-directional LSTM) có khả năng học được mối quan hệ giữa các đặc trưng theo thời gian bằng cách xem xét cả thông tin từ quá khứ và tương lai trong chuỗi dữ liệu. Đồng thời, perceptron một lớp (single layer

perceptron) sẽ tập trung vào việc học các đặc trưng tĩnh không thay đổi theo thời gian. Mô hình này được huấn luyện cùng với nhóm đặc trưng R^8.

Ở mô hình này, vấn đề mất cân bằng lớp được xử lý bằng phương pháp xử lý hàm mất mát focal (focal loss). Phương pháp này giảm ảnh hưởng của mất cân bằng lớp bằng cách điều chỉnh sự chú ý của mô hình đến lớp thiểu số thông qua thiết lập các chỉ số α và γ. Ngoài ra, kỹ thuật Early Stopping được sử dụng cho việc huấn luyện mô hình để dừng huấn luyện khi độ chính xác đã ngừng cải thiện, nhằm tiết kiệm thời gian huấn luyện. Các tham số được sử dụng để xây dựng mô hình LSTM kết hợp SLP được trình bày trong Bảng 4-9.

Bảng 4-9. Tham số của mô hình LSTM kết hợp SLP

Tham số	**Giá trị**
α	0.8
γ	2.0
epochs	2000
batch_size	64

Kết quả xây dựng mô hình

Sau khi lần lượt xây dựng ba mô hình đã đề cập, khả năng dự đoán được đánh giá bằng các chỉ số Accuracy, Precision, Recall, F1 và Log Loss. Bên cạnh đó, do là bài toán dự đoán trên dữ liệu mất cân bằng lớp, nếu chỉ xem các chỉ số Precision và Recall sẽ không phản ánh đúng kết quả dự đoán cho lớp thiểu số. Vì lý do này, chỉ số F1 được sử dụng là chỉ số chính để so sánh các mô hình. Kết quả đánh giá các mô hình được trình bày ở Bảng 4-10.

Bảng 4-10. Kết quả giữa các mô hình MiniRocket-SHAP (R^8), LSTM kết hợp Single Layer Perceptron (R^8) và RF-Static

Model	Accuracy	Precision	Recall	F1-score	Log Loss
MiniRocket-SHAP (R^8)	0.95	0.82	0.77	0.79	1.74
LSTM kết hợp Single Layer Perceptron (R^8)	0.94	0.74	0.80	0.76	1.95

RF-Static	0.93	0.70	0.77	0.73	2.5

4.2. Đánh giá kết quả dự đoán

Sau khi hoàn thành xây dựng mô hình, đồng thời dựa trên các chỉ số đánh giá, mô hình MiniRocket-SHAP với nhóm đặc trưng R^8 đã cho thấy hiệu quả tốt nhất. Mô hình này đạt được độ chính xác là 0.95, Precision là 0.82, Recall là 0.77, F1 là 0.79 và Log Loss là 1.74. Kết quả này xác nhận rằng việc áp dụng mô hình phân loại chuỗi thời gian này để dự đoán khách hàng rời bỏ là rất hiệu quả. Ngoài ra, so sánh kết quả của mô hình MiniRocket-SHAP khi huấn luyện với các nhóm đặc trưng R^8, E^{85}, D^{16} đã chỉ ra rằng nhóm đặc trưng R^8 được chọn lọc và trích xuất đã giúp cải thiện đáng kể kết quả phân loại so với hai nhóm dữ liệu còn lại. Cụ thể, mô hình MiniRocket-SHAP với nhóm đặc trưng E^{85} có chỉ số F1 chỉ là 0.58. Mô hình MiniRocket-SHAP với nhóm đặc trưng D^{16} có độ chính xác là 0.65, Precision là 0.66, Recall là 0.54, F1 Score là 0.48 và Log Loss là 12.5. Những kết quả này cho thấy rõ ràng phương pháp trích xuất và lựa chọn đặc trưng dựa trên giá trị Shapley của mô hình MiniRocket-SHAP đã giúp hiệu suất của mô hình phân loại chuỗi thời gian cải thiện đáng kể.

So sánh MiniRocket-SHAP với các mô hình dự đoán rời bỏ khác cũng đã chứng minh tính hiệu quả của MiniRocket-SHAP. Mô hình sử dụng dữ liệu tĩnh RF-Static có độ chính xác là 0.93, Precision là 0.70, Recall là 0.77, F1 Score là 0.73 và Log Loss là 2.5, thấp hơn so với MiniRocket-SHAP. Tương tự, MiniRocket-SHAP cũng phân loại tốt hơn mô hình LSTM kết hợp SLP, với chỉ số F1 là 0.79 so với 0.76 của mô hình LSTM kết hợp Single Layer Perceptron. Đồng thời, mô hình LSTM kết hợp Single Layer Perceptron cũng đã cho thấy khả năng kết hợp tốt giữa dữ liệu thời gian và dữ liệu tĩnh không đổi theo thời gian.

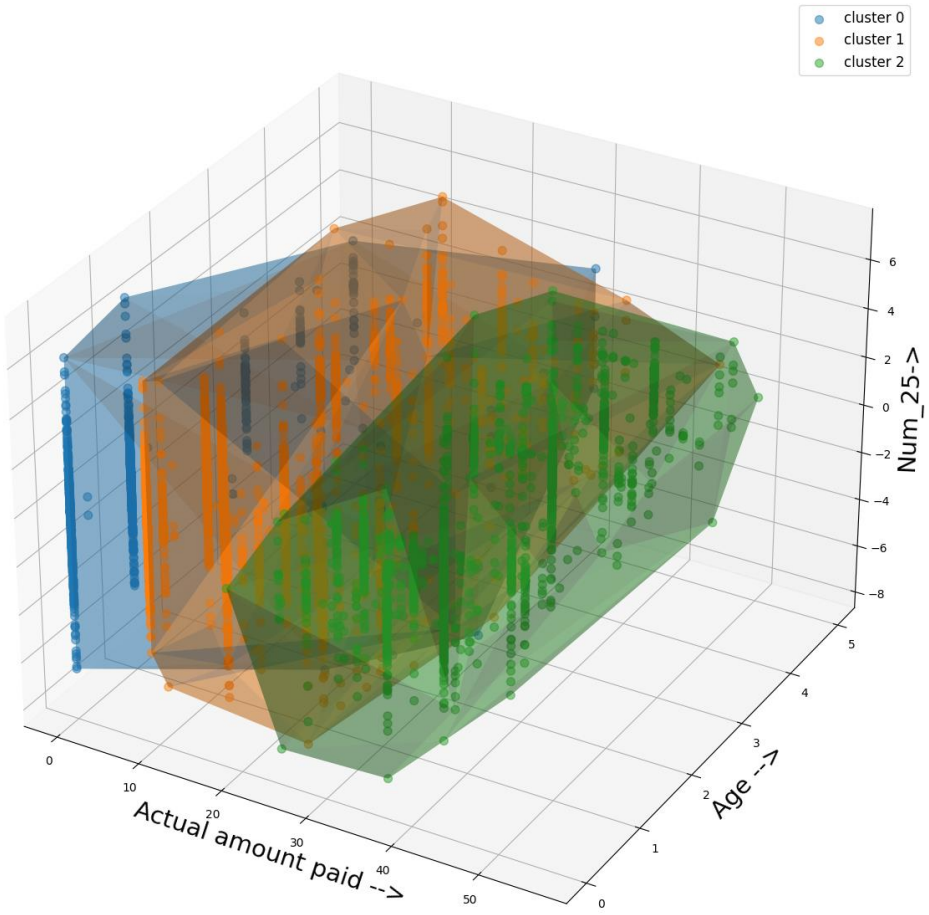
4.3. Phân tích mở rộng

4.3.1. Xác định nguyên nhân rời bỏ

Ngoài đưa ra kết quả dự đoán rời bỏ, mô hình MiniRocket-SHAP còn mở ra khả năng phân tích chuyên sâu hơn giúp đề xuất chiến lược giữ chân khách hàng cho doanh nghiệp. Nhóm nghiên cứu đề xuất sử dụng độ quan trọng đặc trưng từ giá trị Shapley cho việc xác định yếu tố ảnh hưởng đến quyết định rời bỏ. Ngoài ra, kỹ thuật phân cụm

chuỗi thời gian cũng được đề xuất áp dụng cho các khách hàng được dự đoán rời bỏ nhằm tìm ra đặc tính của các khách hàng này.

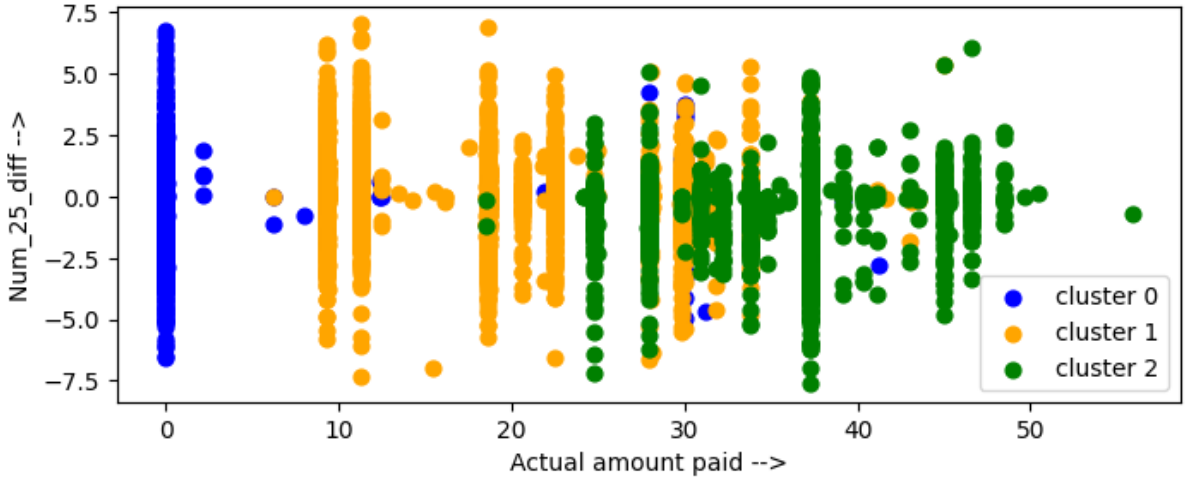
Từ Hình 4-8, có thể thấy rằng trong số nhiều đặc trưng được sử dụng để xây dựng mô hình dự đoán, đặc trưng "actual_amount_paid", hay có thể diễn giải là số tiền thực tế phải chi trả cho gói đăng ký, là đặc trưng tác động nhiều nhất đến kết quả phân lớp. Điều này lý giải rằng quyết định rời bỏ của khách hàng có nguyên nhân đến từ số tiền thực tế khách hàng phải chi trả. Đặc trưng có mức độ ảnh hưởng lớn tiếp theo là "age_group", hay được hiểu là nhóm tuổi, ngoài ra các đặc trưng còn lại có sự tác động nhỏ nhưng không đáng kể khi so sánh giá trị Shapley với hai đặc trưng vừa nêu.



Hình 4-10. Biểu đồ 3D thể hiện các cụm khách hàng rời bỏ

Nhằm mục đích hiểu hơn về nguyên nhân rời bỏ của khách hàng, kỹ thuật phân cụm chuỗi thời gian Time Series K-means được sử dụng để phân cụm các khách hàng được dự đoán là sẽ rời bỏ. Thuật toán phân cụm giúp tìm ra các cụm khách hàng có đặc

tính giống nhau, dựa trên trên ba đặc trưng quan trọng nhất là “actual_amount_paid”, “age_group” và “num_25_diff_mean”. Đặc tính của khách hàng rời bỏ có thể được nhận xét bằng cách quan sát phân phối của các mẫu dữ liệu trên không gian các chiều dữ liệu. Quan sát hình 4-10 cho thấy sự khác biệt giữa ba cụm khách hàng chủ yếu theo chiều của biến “actual_amount_paid”, tức là số tiền thực tế đã chi trả.



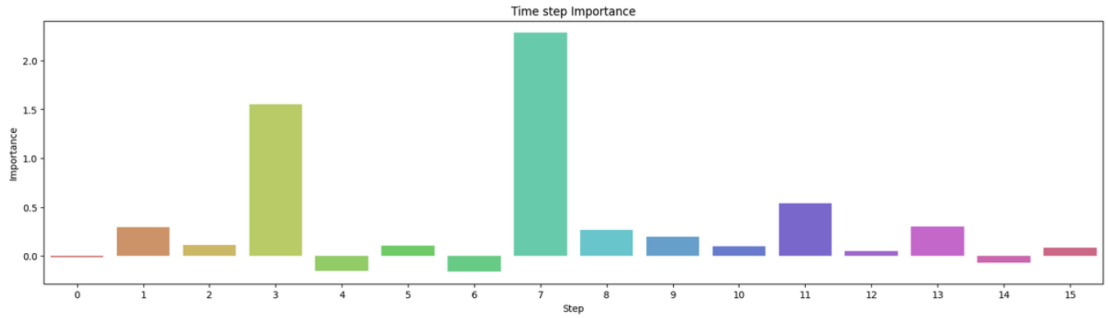
Hình 4-11. Biểu đồ 2D thể hiện các cụm khách hàng rời bỏ

Biểu đồ phân tán hai chiều ở Hình 4-11 cho thấy một góc nhìn rõ ràng hơn về xu hướng phân bố không khác biệt theo chiều của biến “num_25_diff” mà chỉ phân bố khác biệt theo chiều của biến “actual_amount_paid”. Kết quả phân cụm chuỗi thời gian này dẫn đến kết luận rằng “số tiền khách hàng phải chi trả” là yếu tố ảnh hưởng nhiều nhất đến quyết định rời bỏ của khách hàng. Hai yếu tố còn lại (nhóm tuổi và giá trị trung bình chênh lệch 25 ngày) có ảnh hưởng nhất định đến việc phân chia cụm, nhưng mức độ ảnh hưởng không cao bằng yếu tố “số tiền thực tế đã trả”. Ý nghĩa của ba cụm khách hàng được giải thích như sau:

- Cụm 0 - 1024 khách hàng: Nhóm khách hàng rời bỏ phải chi trả ít cho gói đăng ký.
- Cụm 1 - 5128 khách hàng: Nhóm khách hàng rời bỏ phải chi trả trung bình cho gói đăng ký.
- Cụm 2 - 2063 khách hàng: Nhóm khách hàng rời bỏ phải chi trả cao cho gói đăng ký.

Khi so sánh số lượng khách hàng ở mỗi cụm, có thể nhận thấy có 1024 khách hàng ở cụm 0, còn gọi là cụm các khách hàng chi trả ít. Số lượng khách hàng ở cụm chi trả trung bình là gấp 5 lần và là gấp 2 lần ở cụm chi trả cao. Điều này biểu thị cho xu hướng khách hàng ít rời bỏ hơn khi chỉ phải chi trả chi phí cho gói đăng ký ở mức thấp. Từ đó, có được đề xuất doanh nghiệp đưa ra chiến lược giữ chân khách hàng bằng cách giảm số tiền thực tế khách hàng phải chi trả, thông qua khuyến mãi hoặc giảm các loại thuế.

4.3.2. Thời điểm ảnh hưởng đến quyết định rời bỏ



Hình 4-12. Time step importance

Ngoài độ quan trọng của đặc trưng, mô hình MiniRocket-SHAP còn giúp thu được độ quan trọng của bước thời gian trong chuỗi thời gian. Kết quả độ quan trọng bước thời gian ở Hình 4-12 cho thấy bước thời gian thứ 7 và thứ 3 trong khoảng n ϵ [0;16] là hai thời điểm có tác động nhiều nhất để mô hình dự đoán liệu khách hàng sẽ rời bỏ hay không. Điều này cho thấy các dấu hiệu của hành vi rời bỏ khách hàng đã xuất hiện từ 8 tuần trước khi khách hàng thực sự rời bỏ, với tuần cuối cùng trong chuỗi thời gian là thời điểm khách hàng rời bỏ. Khi xác định được thời điểm gây ảnh hưởng nhiều nhất đến quyết định rời bỏ, doanh nghiệp nên đưa ra hai chiến lược sau để làm tăng tỷ lệ giữ chân khách hàng: (1) Thực hiện phân tích chẩn đoán (diagnostic analysis) dựa trên dữ liệu ở tuần đó để tìm ra nguyên nhân cụ thể, xác định sự kiện nào làm cho khách hàng quyết định rời bỏ ở thời điểm này; và (2) Dựa vào nguyên nhân rời bỏ đã xác định, sử dụng chiến dịch giữ chân khách hàng để can thiệp quyết định rời bỏ sớm nhất có thể tính từ thời điểm đó.

Việc sử dụng kết quả độ quan trọng đặc trưng và bước thời gian có được từ mô hình MiniRocket-SHAP kết hợp phân cụm chuỗi thời gian Time Series K-means giúp thu được những hiểu biết quan trọng về quyết định rời bỏ của khách hàng. Điều này nhấn mạnh tiềm năng và tính ứng dụng cao của mô hình MiniRocket-SHAP trong việc dự đoán và can thiệp quyết định rời bỏ của khách hàng, giúp doanh nghiệp giữ chân khách hàng và tối ưu lợi nhuận.

Tóm tắt chương 4

Trong chương 4, nhóm đưa ra kết quả của từng bước theo mô hình đề xuất ở chương 3 bao gồm các kết quả, cơ sở lý luận và nhận xét liên quan. Đồng thời, sau khi kết thúc thực nghiệm, nhóm nghiên cứu tiến hành đánh giá chi tiết về kết quả nghiên cứu.

CHƯƠNG 5. THẢO LUẬN VÀ ĐỀ XUẤT

5.1. Kết quả tổng thể của nghiên cứu

Sau khi khảo lược và tìm ra các vấn đề cần được giải quyết trong bài toán dự đoán rời bỏ, giải pháp được xác định là một mô hình dự đoán rời bỏ trên dữ liệu chuỗi thời gian để không bỏ qua sự thay đổi trong hành vi khách hàng, đồng thời giúp giải thích được nguyên nhân và thời điểm khách hàng quyết định rời bỏ. Do đó, nghiên cứu này định hướng tập trung thiết kế và đề xuất mô hình phân loại chuỗi thời gian MiniRocket-SHAP trong dự đoán sự rời bỏ. Khi so sánh với các mô hình khác như RF-Static và LSTM kết hợp Single Layer Perceptron trên bộ dữ liệu kéo dài trong 4 tháng, MiniRocket-SHAP đã thể hiện hiệu quả cao hơn trong dự đoán rời bỏ so với các mô hình truyền thống và phân loại chuỗi thời gian khác. Điều này mở ra tiềm năng lớn cho việc áp dụng mô hình này để dự đoán sớm sự rời bỏ, hỗ trợ cho các chiến dịch giữ chân khách hàng.

Giá trị Shapley thu được từ mô hình MiniRocket-SHAP cũng mở ra nhiều hướng phân tích sâu hơn về hành vi rời bỏ của khách hàng, làm nổi bật tính ứng dụng và tiềm năng của phương pháp này trong thực tiễn. Bài nghiên cứu đã trình bày phương pháp, cách thức ứng dụng mô hình MiniRocket-SHAP để giải thích mức độ ảnh hưởng của các đặc trưng đối với kết quả dự đoán. Nhờ vậy, nguyên nhân rời bỏ có thể được xác định chỉ sau vài bước phân tích mở rộng. Đồng thời, mô hình MiniRocket-SHAP cho phép xác định thời điểm ảnh hưởng nhiều nhất đến quyết định rời bỏ của khách hàng. Phương pháp phân tích mới mẻ này giúp giải thích được các sự kiện ảnh hưởng đến quyết định rời bỏ đã xảy ra lúc nào, cũng như giúp xác định thời điểm can thiệp quyết định rời bỏ sớm và hiệu quả nhất. Kết quả này cho thấy mô hình MiniRocket-SHAP là lựa chọn phù hợp cho dự đoán sự rời bỏ sớm, giúp các doanh nghiệp can thiệp vào hành vi rời bỏ kịp thời.

Ngoài ra, Mô hình LSTM kết hợp Single Layer Perceptron cũng đã chứng minh tính hiệu quả của việc kết hợp dữ liệu thời gian và dữ liệu tĩnh như dữ liệu nhân khẩu học làm đầu vào cho phân loại chuỗi thời gian. Single Layer Perceptron đã đóng vai trò nhận diện đặc tính của các đặc trưng không thay đổi theo thời gian, giúp cải thiện hiệu quả dự đoán.

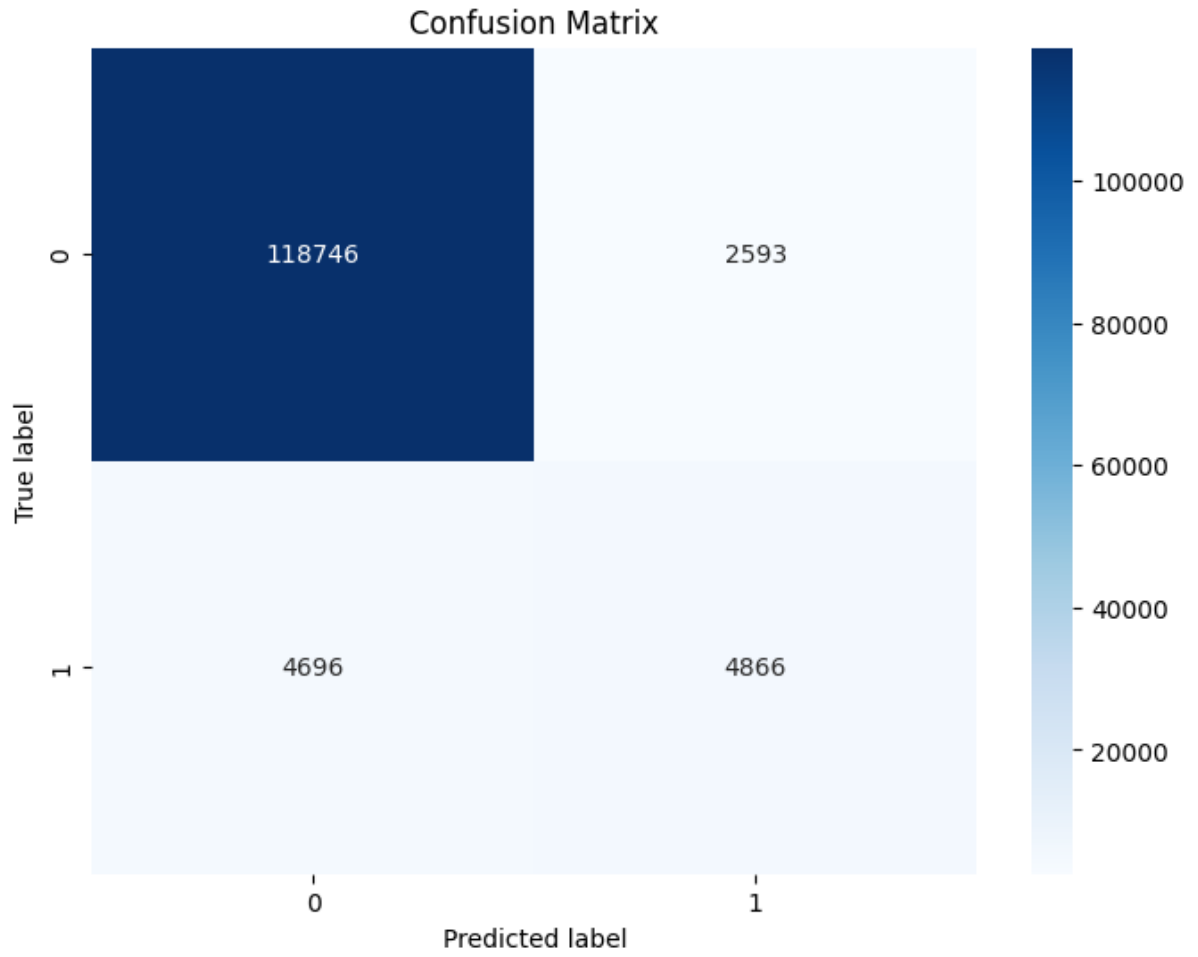
5.2. Thảo luận

Với kết quả dự đoán rời bỏ đạt được, có thể thấy tất cả mô hình được thực nghiệm đều chưa tối ưu với dữ liệu mất cân bằng lớp, với điểm F1 của mô hình cao nhất là 0.79. Một số nghiên cứu khác cũng sử dụng phân lớp chuỗi thời gian linh hoạt và đưa ra kết quả rất tích cực. Alboukaey và cộng sự (2020) đã đề xuất hướng tiếp cận dự đoán rời bỏ ở cấp độ ngày, với ba mô hình RFM-based, CNN-based và LSTM-based được đề xuất. Óskarsdóttir el al. (2018) cũng đã đề xuất mô hình dự đoán rời bỏ ở cấp độ hàng tuần. Cả hai bài nghiên cứu này đều sử dụng dữ liệu trong ngành viễn thông, với tỉ lệ giữa hai lớp khá cân bằng. Tuy nhiên, Alboukaey và cộng sự (2020) đạt được điểm F1 là 0.571 ở mô hình tốt nhất là LSTM-based. Dù chưa tương đồng hoàn toàn về dữ liệu và môi trường thực nghiệm, kết quả này cho thấy mô hình MiniRocket-SHAP là tốt hơn so với các nghiên cứu trước với thiết lập thực nghiệm gần giống.

Ngoài ra mô hình MiniRocket-SHAP được đề xuất có những điểm khác biệt và cải thiện đáng kể so với các nghiên cứu trước nêu trên. Đầu tiên, mô hình MiniRocket-SHAP là giải thích được (interpretable). Ngoài việc giải thích độ quan trọng đặc trưng, nó còn giúp giải thích được thời điểm nào ảnh hưởng đến quyết định rời bỏ thông qua độ quan trọng của bước thời gian. Thứ hai, nhóm nghiên cứu đề xuất sử dụng dữ liệu chuỗi thời gian theo tuần. Việc sử dụng chuỗi thời gian theo tuần không chỉ giúp giảm tài nguyên tính toán so với chuỗi thời gian theo ngày (giảm 7 lần), mà còn đảm bảo được độ chính xác của mô hình dự đoán. Thứ ba, việc lấy mẫu để thiết lập cho thực nghiệm được nhóm nghiên cứu thực hiện bám sát thực tế hơn so với Óskarsdóttir el al. (2018). Nhóm nghiên cứu đã chọn dữ liệu thực nghiệm dựa trên 4 cửa sổ thời gian cách nhau 1 tuần, trong đó cửa sổ thời gian cuối cùng là dữ liệu kiểm thử (test). Phương pháp lấy mẫu mày bám sát với thực tế hơn so với việc lấy mẫu tập huấn luyện và kiểm tra trên chung một khung thời gian như nghiên cứu trước đó. Cuối cùng, các nghiên cứu trước cũng chưa chú trọng vào bước xử lý và trích xuất đặc trưng các đặc trưng mới. Kết quả so sánh các không gian đặc trưng cho thấy việc trích xuất đặc trưng mới là chìa khóa để đạt hiệu quả mô hình cao.

Một vấn đề khác cần được đề cập là việc đánh giá chi phí của các dự đoán sai đóng vai trò quan trọng để tận dụng tối đa giá trị của mô hình dự đoán rời bỏ. Ma trận

được trình bày ở Hình 5-1 cho thấy tồn tại các dự đoán sai. Cần dựa trên bối cảnh bài toán để xác định rõ ràng chi phí của các dự đoán False Negative và False Positive nhằm xác định đúng rủi ro mà vấn đề mất cân bằng lớp gây ra. Khi xác định được TPR (tỉ lệ true positive) và FPR (tỉ lệ false positive), việc giảm rủi ro và chi phí của việc dự đoán sai có thể đạt được thông qua điều chỉnh mô hình để thay đổi hai tỷ lệ này.



Hình 5-1. Confusion Matrix trên mô hình đề xuất

5.3. Đề xuất sử dụng mô hình

Từ kết quả thực nghiệm, nhóm nghiên cứu đề xuất sử dụng mô hình tốt nhất là MiniRocket-SHAP cho việc dự đoán rời bỏ cũng như xác định nguyên nhân, thời điểm rời bỏ. Như đã được đề cập, phương pháp tiếp cận được sử dụng dựa trên dữ liệu chuỗi thời gian theo tuần. Tuy nhiên, phương pháp này nhấn mạnh tính linh hoạt, đồng nghĩa với việc có thể được áp dụng bất kỳ thời điểm nào để phục vụ cho các chiến dịch giữ chân khách hàng. Dù vậy, nhóm nghiên cứu đề xuất sử dụng mô hình phân lớp chuỗi

thời gian thường xuyên, cụ thể là các phân tích dự đoán nên được thực hiện cách nhau mỗi tuần hoặc không quá một tháng. Lý do được đưa ra là hành vi của khách hàng dễ bị thay đổi theo các tuần trong tháng và trong thời gian dẫn đến quyết định rời bỏ khách hàng, họ bắt đầu hành xử khác đi. Alboukaey và cộng sự (2020) đã đưa ra một ví dụ cho vấn đề này như sau: giả sử x_monthly = (600, 600, 600) là một vectơ biểu diễn chi tiêu hàng tháng của một khách hàng trong ba tháng gần đây. Theo cách biểu diễn này, khách hàng này là một khách hàng không rời bỏ dịch vụ có hành vi ổn định. Tuy nhiên, nó cũng có thể biểu diễn cho một khách hàng rời bỏ dịch vụ với chi tiêu ổn định trong 8 tuần đầu tiên (tức là hai tháng đầu) và sau đó có một đỉnh điểm theo sau bởi một đường dốc vào đầu tháng cuối, ví dụ x_weekly = (20, 20, 20, 20, 180, 160, 120, 100, 40, 0, 0, 0, 0). Do đó, nếu chỉ thực hiện dự đoán mỗi tháng và sử dụng dữ liệu được trình bày theo tháng, các chiến dịch giữ chân khách hàng sẽ không kịp phản ứng nếu xuất hiện một đường dốc như vậy. Để đảm bảo tỷ lệ giữ chân khách hàng, việc thực hiện dự đoán rời bỏ thường xuyên, cách nhau không quá một tháng là rất quan trọng.

Tóm tắt chương 5

Trong chương 5, nhóm nghiên cứu thảo luận về kết quả nghiên cứu, tiềm năng của phương pháp cũng như so sánh với các hướng nghiên cứu khác. Các hướng phân tích khai thác kết quả nghiên cứu, các kỹ thuật được đề xuất sử dụng mở rộng sử dụng kết quả dự đoán cũng được trình bày chi tiết.

CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận

Tóm lại, mục tiêu bài nghiên cứu này nhằm dự đoán khách hàng rời bỏ trong lĩnh vực kinh doanh dịch vụ trực tuyến, sử dụng bộ dữ liệu từ nền tảng âm nhạc trực tuyến KKBOX, sử dụng phân loại chuỗi thời gian thể hiện sự thay đổi hành vi của khách hàng. Về đóng góp trên khía cạnh phương pháp luận, nhóm trình bày một cách tiếp cận mới để trích xuất dữ liệu chuỗi thời gian đa biến từ các bản ghi hành vi nghe nhạc của khách hàng.

Trên khía cạnh kinh doanh, bài nghiên cứu đã phân tích những yêu cầu kinh doanh của các doanh nghiệp trong lĩnh vực kinh doanh dịch vụ, trong có có các doanh nghiệp cung cấp nền tảng âm nhạc trực tuyến, và thành công xây dựng một mô hình mang tính thực tế, chú trọng vào phân tích hành vi khách hàng theo tuần. Dựa trên khảo sát và phân tích các vấn đề cần giải quyết, nhóm nghiên cứu đã xây dựng một mô hình dự đoán rời bỏ trên dữ liệu chuỗi thời gian MiniRocket-SHAP. Mô hình này không chỉ giúp dự đoán sự rời bỏ mà còn giúp giải thích nguyên nhân và thời điểm quyết định của khách hàng. So sánh với các mô hình truyền thống khác, MiniRocket-SHAP đã thể hiện hiệu quả cao hơn trong việc dự đoán sự rời bỏ. Giá trị Shapley từ mô hình cũng giúp chúng tôi hiểu sâu hơn về hành vi của khách hàng và xác định ảnh hưởng của các đặc trưng đối với kết quả dự đoán. Trong đó, điểm mới nổi bật nhất là khả năng xác định bước thời gian, hay thời điểm ảnh hưởng nhiều nhất đến quyết định rời bỏ của khách hàng.

Với những điểm mới này, mô hình mang lại một đóng góp quan trọng cho các doanh nghiệp, giúp dự đoán hành vi khách hàng rời bỏ chính xác và kịp thời hơn. Đây là cơ sở then chốt để đề xuất các chiến lược giữ chân khách hàng hiệu quả, góp phần cho sự phát triển lâu dài của doanh nghiệp.

6.2. Hạn chế và hướng phát triển

Mặc dù đã thành công trong việc triển khai mô hình phân lớp chuỗi thời gian linh hoạt giúp dự đoán rời bỏ sớm, nghiên cứu vẫn tồn tại một số điểm hạn chế. Việc thực nghiệm chỉ thực hiện phân tích tích dựa trên dữ liệu về khách hàng nghe nhạc trên

một nền tảng âm nhạc, cụ thể là KKBOX và chưa được thực hiện rộng rãi trên nhiều bộ dữ liệu khác nhau trong cùng ngành và số mẫu dữ liệu được sử dụng chưa lớn. Đồng thời, vấn đề giới hạn về không gian siêu tham số khi điều chỉnh mô hình bắt nguồn từ hạn chế tài nguyên máy tính gây hạn chế cho kết quả dự đoán. Giới hạn về tài nguyên máy tính cũng hạn chế sự so sánh phương pháp chọn đặc trưng sử dụng mô hình MiniRocket-SHAP với các phương pháp khác, trong đó có Exhaustive Feature Selection được đánh giá mang lại hiệu quả cao bởi Schadl và cộng sự (2017). Sự hạn chế này cũng xuất phát từ việc có rất ít, hay không có mô hình tối ưu cụ thể với bài toán này, được đề xuất trước đây giúp chọn đặc trưng và giải thích mô hình cho bài toán phân lớp chuỗi thời gian.

Với những kết quả đạt được và một số những hạn chế của nghiên cứu hiện tại, nhóm nhận thấy những hướng đến mở rộng và phát triển trong tương lai. Việc thực nghiệm được thực hiện trên bộ dữ liệu lớn hơn và đa dạng hơn về các đặc trưng. Để đánh giá hiệu quả mô hình, nhóm sẽ tiến hành so sánh phương pháp chọn đặc trưng kết hợp giữa MiniRocket và SHAP với phương pháp Exhaustive Feature Selection, sau đó so sánh kết quả với các phương pháp phân lớp chuỗi thời gian khác như CNN, HIVE-COTE,... để đánh giá một cách toàn diện và hoàn thiện mô hình. Ngoài ra, đối với dữ liệu chuỗi thời gian, các kỹ thuật xử lý mất cân bằng lớp phổ biến như SMOTE, Borderline-SMOTE có thể sẽ không hiệu quả, do chúng không xem xét đặc điểm thời gian (tức là các đặc điểm thay đổi dần dần và liên tục) của dữ liệu chuỗi thời gian (Zhao và cộng sư, 2022). Do đó, một hướng phát triển cần thiết là giải quyết vấn đề này bằng cách áp dụng các kỹ thuật xử lý mất cân bằng lớp được tinh chỉnh cho dữ liệu chuỗi thời gian.

Tóm tắt chương 6

Trong chương 6, từ những kết quả thực nghiệm, nhóm rút ra được những đóng góp chính cho các doanh nghiệp trong việc dự đoán khách hàng rời bỏ. Song vẫn còn một số hạn chế, và sẽ được khắc phục và mở rộng trong các nghiên cứu sau này.

DANH MỤC TÀI LIỆU THAM KHẢO

Addison Howard, Arden Chiu, Mark McDonald, msla, Wendy Kan, & Yianchen. (2017). *WSDM - KKBox's Music Recommendation Challenge | Kaggle*. https://www.kaggle.com/c/kkbox-music-recommendation-challenge/overview/prizes

Aggarwal, A., Kasiviswanathan, S. P., Xu, Z., Feyisetan, O., & Teissier, N. (2021). Label Inference Attacks from Log-loss Scores. *Proceedings of Machine Learning Research*, *139*.

Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic behavior based churn prediction in mobile telecom. *Expert Systems with Applications*, *162*. https://doi.org/10.1016/j.eswa.2020.113779

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1). https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. https://doi.org/10.1201/9781315139470

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*(3 PART 1). https://doi.org/10.1016/j.eswa.2008.05.027

Cabello, N., Naghizade, E., Qi, J., & Kulik, L. (2020). Fast and accurate time series classification through supervised interval search. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2020-November*. https://doi.org/10.1109/ICDM50108.2020.00107

Chandar, M., Arijit, L., & Krishna, P. (2006). Modeling churn behavior of bank customers using predictive data mining tech. *National Conference on Soft Computing Techniques for Engineering Applications (SCT-2006)*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*. https://doi.org/10.1613/jair.953

Chen, Y., Xie, X., Lin, S. De, & Chiu, A. (2018). WSDM Cup 2018: Music Recommendation and churn prediction. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining, 2018-Febuary*. https://doi.org/10.1145/3159652.3160605

Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3447548.3467231

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys*, *45*(1). https://doi.org/10.1145/2379776.2379788

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. In *International Journal of Data Science and Analytics* (Vol. 14, Issue 3). https://doi.org/10.1007/s41060-022-00312-5

Guilleme, M., Masson, V., Roze, L., & Termier, A. (2019). Agnostic local explanation for time series classification. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2019-November*. https://doi.org/10.1109/ICTAI.2019.00067

Hegde, S. K., Hegde, R., Nanda, S. S., Phatak, G., Hongal, P., & Gowda, V. D. (2023). Customer Churn Analysis in Financial Domain using Deep Intelligence Network. *IDCIoT 2023 - International Conference on Intelligent Data Communication Technologies and Internet of Things, Proceedings*. https://doi.org/10.1109/IDCIoT56793.2023.10053473

Hochreiter, S., & Schmidhuber, J. (1997). Long Short Term Memory. Neural Computation. *Neural Computation*, *9*(8).

Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, *367–368*. https://doi.org/10.1016/j.ins.2016.05.040

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K.,

Rätsch, G., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, *26*(3). https://doi.org/10.1038/s41591-020-0789-4

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4). https://doi.org/10.1007/s10618-019-00619-1

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P. A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, *34*(6). https://doi.org/10.1007/s10618-020-00710-y

Jeyakumar, J. V., Noor, J., Cheng, Y. H., Garcia, L., & Srivastava, M. (2020). How can I explain this to you? An empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, *2020-December*.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Kumar, S., & D., C. (2016). A Survey on Customer Churn Prediction using Machine Learning Techniques. *International Journal of Computer Applications*, *154*(10). https://doi.org/10.5120/ijca2016912237

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2). https://doi.org/10.1007/s00607-021-00908-y

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11). https://doi.org/10.1109/5.726791

Liu, H. Y., Gao, Z. Z., Wang, Z. H., & Deng, Y. H. (2022). Time Series Classification with Shapelet and Canonical Features. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178685

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*.

Meenal, R., Michael, P. A., Pamela, D., & Rajasekaran, E. (2021). Weather prediction using random forest machine learning model. *Indonesian Journal of Electrical Engineering and Computer Science*, *22*(2). https://doi.org/10.11591/ijeecs.v22.i2.pp1208-1215

Neslin, S. A., Gupta, S., Kamakura, W., Junxiang, L. U., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2). https://doi.org/10.1509/jmkr.43.2.204

Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., & Vanthienen, J. (2018). Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Systems with Applications*, *106*. https://doi.org/10.1016/j.eswa.2018.04.003

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, *96*(1). https://doi.org/10.1080/00220670209598786

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLoS ONE*, *18*(12 December). https://doi.org/10.1371/journal.pone.0289724

Pustokhina, I. V., Pustokhin, D. A., RH, A., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. *Information Processing and Management*, *58*(6). https://doi.org/10.1016/j.ipm.2021.102706

Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context. *International Journal of Electrical and Computer Engineering*, *8*(4). https://doi.org/10.11591/ijece.v8i4.pp2367-2383

Rosenberg, L. J., & Czepiel, J. A. (1984). A marketing approach for customer retention. In *Journal of Consumer Marketing* (Vol. 1, Issue 2). https://doi.org/10.1108/eb008094

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6). https://doi.org/10.1037/h0042519

Schadl, K., Vassar, R., Cahill-Rowley, K., Yeom, K. W., Stevenson, D. K., & Rose, J. (2018). Prediction of cognitive and motor development in preterm children using exhaustive feature selection and cross-validation of near-term white matter microstructure. *NeuroImage: Clinical*, *17*. https://doi.org/10.1016/j.nicl.2017.11.023

Schäfer, P. (2016). Scalable time series classification. *Data Mining and Knowledge Discovery*, *30*(5). https://doi.org/10.1007/s10618-015-0441-y

Tan, C. W., Dempster, A., Bergmeir, C., & Webb, G. I. (2022). MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, *36*(5). https://doi.org/10.1007/s10618-022-00844-1

Theissler, A., Spinnato, F., Schlegel, U., & Guidotti, R. (2022). Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. *IEEE Access*, *10*. https://doi.org/10.1109/ACCESS.2022.3207765

Tianyuan, Z., & Moro, S. (2021). Research Trends in Customer Churn Prediction: A Data Mining Approach. *Advances in Intelligent Systems and Computing*, *1365 AIST*. https://doi.org/10.1007/978-3-030-72657-7_22

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10). https://doi.org/10.1016/j.eswa.2009.05.032

Venkatesh, B., & Anuradha, J. (2019). A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, *19*(1). https://doi.org/10.2478/CAIT-2019-0001

Verbraken, T., Verbeke, W., & Baesens, B. (2014). Profit optimizing customer churn prediction with Bayesian network classifiers. *Intelligent Data Analysis*, *18*(1). https://doi.org/10.3233/IDA-130625

Wang, W. K., Chen, I., Hershkovich, L., Yang, J., Shetty, A., Singh, G., Jiang, Y., Kotla, A., Shang, J. Z., Yerrabelli, R., Roghanizad, A. R., Shandhi, M. M. H., & Dunn, J. (2022). A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications. In *Sensors* (Vol. 22, Issue 20). https://doi.org/10.3390/s22208016

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(1). https://doi.org/10.1007/s11229-022-03485-5

Zeithaml, V. A. (2000). Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science*, *28*(1). https://doi.org/10.1177/0092070300281007