TẠP CHÍ KHOA HỌC ĐẠI HỌC HẠ LONG
Scientific Journal of Ha Long University
http://uhl.edu.vn/

ÁP DỤNG MÔ HÌNH MÁY HỌC CHO BÀI TOÁN DỰ ĐOÁN HÀNH VI THUÊ BAO TRONG VIỄN THÔNG

Nguyễn Văn Hậu[1*], Phạm Vũ Văn[2], Đàm Quang Thịnh[1], Nguyễn Minh Huyền[4], Nguyễn Đức Tuấn Anh[1], Nguyễn Mậu Trường Giang[3]

[1] Khoa Công nghệ Thông tin, Đại học Sư phạm Kỹ thuật Hưng Yên
[2] TT Công nghệ Thông tin, VNPT Hải Dương
[3] Trung tâm Sáng tạo, Công ty Công nghệ thông tin VNPT
[4] Viện Kinh tế và Kinh doanh Quốc tế, Đại học Ngoại thương Hà Nội
[*] Email: nvhau66@gmail.com

Ngày nhận bài: 26/9/2021 *Ngày chấp nhận đăng: 15/12/2021*

TÓM TẮT

Xu hướng hiện nay, nghiên cứu và áp dụng các mô hình máy học vào các lĩnh vực trong cuộc sống đang được quan tâm bởi cả hai cộng đồng: nhà khoa học và doanh nghiệp. Trong bài báo này, chúng tôi tiến hành tìm hiểu và áp dụng mô hình máy học cho lĩnh vực viễn thông. Cụ thể, chúng tôi sử dụng các mô hình phân lớp để dự đoán hành vi của thuê bao. Chúng tôi thực hiện đánh giá trên 10 mô hình máy học khác nhau dựa trên các giá trị đánh giá F1-Score, Precision, Recall, và Accuracy. Đánh giá được thực hiện dựa trên 6000 mẫu tập tin khác nhau được thu thập từ Trung tâm Công nghệ Thông tin, VNPT Hải Dương. Thực nghiệm chứng tỏ rằng sử dụng các mô hình học máy cho bài toán là một cách tiếp cận phù hợp. Đặc biệt, mô hình Gradient Boosting cho độ hồi tưởng (recall) rất cao – 0.986. Đây thực sự là một kết quả ấn tượng và là tiền đề để cải tiến, phát triển các mô hình mới đạt hiệu quả cao hơn trong tương lai nhằm giúp doanh nghiệp có thể đưa ra những quyết định kinh doanh phù hợp hơn với từng loại khách hàng.

Từ khóa: *hành vi thuê bao khách hàng, khai phá dữ liệu, máy học, mô hình phân lớp, phân lớp*

USING MACHINE LEARNING MODELS TO PREDICT THE CONSUMER BEHAVIOR IN TELECOMMUNICATION

ABSTRACT

Researching and applying machine learning models in real-life domains tend to attract communities such as scientists and businesses. This paper describes the results of our analysis of applying machine learning models to the field of telecommunications. Specifically, in our research, we used classification models to predict consumer behavior and performed evaluation on 10 different machine learning models based on F1-Score, Precision, Recall, and Accuracy. Experiments have been carried out on 6000 different samples collected from the Center of Information Technology, VNPT Hai Duong. We found that using machine learning models for the problem is a very promising approach to predict consumer behaviors in telecommunication. In particular, the Gradient Boosting model provides a very high recall – 0.986. This significant result is the premise to improve and develop new models with higher efficiency in the future for businesses to make more suitable business decisions for each type of customers.

Keywords: *classification, consumer behavior, data mining, machine learning, telecom*

1. ĐẶT VẤN ĐỀ

Hiện nay, các doanh nghiệp viễn thông trong nước và quốc tế đang từng bước chuyển mình theo xu thế mới, giai đoạn mới đó là giai đoạn công nghệ số, dịch vụ số. Ở giai đoạn này, song song phát triển các nhóm dịch vụ số mới vẫn tiếp tục tăng cường cạnh tranh, củng cố thị phần các dịch vụ viễn thông truyền thống và đặc biệt chú trọng tăng cường ứng dụng giải pháp công nghệ thông tin để nâng cao chất lượng các chương trình khuyến mại, chăm sóc khách hàng và hạ tầng dịch vụ.

Khối dữ liệu khổng lồ của các nhà cung cấp dịch vụ bao gồm những thông tin từ nhân khẩu học đến thông tin hành vi sử dụng dịch vụ viễn thông, đều được lưu trữ một cách chi tiết nhằm phục vụ các hoạt động kinh doanh và vận hành hàng ngày. Nếu dữ liệu đó được khai phá, chúng ta sẽ có được tri thức quý báu về thị trường, khách hàng, sản phẩm để giúp tổ chức có quyết định phù hợp đối với quá trình kinh doanh và hoạch định chiến lược phát triển. Sử dụng các kỹ thuật phân tích, khai phá dữ liệu (KPDL) cùng các các thuật toán và mô hình máy học phù hợp cho phép các nhà mạng đưa ra các quyết định thông minh, hiệu quả và kịp thời. Có khá nhiều các bài toán phân tích trong lĩnh vực viễn thông phục vụ cho các hoạt động điều hành và kinh doanh, được chia thành các ba nhóm chủ đề chính sau:

Tiếp thị, bán hàng và quan hệ khách hàng: Dữ liệu của các nhà cung cấp dịch vụ có một lượng lớn dữ liệu hợp đồng, chi tiết cuộc gọi của từng khách hàng, qua đó có cơ hội tạo hồ sơ khách hàng với các điểm số, phân khúc và phân loại khách hàng bằng cách sử dụng nhiều phương pháp và kỹ thuật KPDL. Sau đó, các mẫu và mô hình được khai phá trong dữ liệu đó có thể được sử dụng cho các mục đích tiếp thị, bán hàng và chăm sóc khách hàng. Ví dụ như: phân công và quản lý chiến dịch, cải thiện hiệu quả kênh, phát triển chương trình khách hàng thân thiết, dự đoán doanh số bán sản phẩm và xu hướng của khách hàng, dự đoán và quản lý thời gian, chăm sóc khách hàng tốt hơn. Tất cả những hoạt động này sẽ đóng góp đáng kể vào việc tăng lợi nhuận, cải thiện hiệu quả tài chính và khả năng tương thích, thích ứng với nhu cầu của khách hàng (Kabakchieva, 2009). Một trong những nghiên cứu điển hình của chủ đề này là của Công ty Viễn thông Syriatel: Phân tích dự đoán sử dụng dữ liệu lớn (Big Data) để tăng lòng trung thành của khách hàng (Wassouf & nnk., 2020).

Phát hiện gian lận viễn thông: Gian lận viễn thông được định nghĩa là "bất kỳ hoạt động nào sử dụng dịch vụ viễn thông mà không có ý định trả tiền" (Hilas & Mastorocostas, 2008). Gian lận là một vấn đề nghiêm trọng và nó không chỉ dẫn đến thất thoát doanh thu cho các nhà cung cấp dịch vụ viễn thông mà đôi khi còn là gánh nặng cho khách hàng. Các phương pháp phổ biến nhất được sử dụng để phát hiện gian lận dựa trên phân tích hoạt động của người dùng, so sánh hành vi mới so với hành vi cũ. Còn Weiss (Weiss, 2009) phân loại gian lận viễn thông "gian lận đăng ký: xảy ra khi khách hàng mở tài khoản với ý định không bao giờ trả tiền và gian lận chồng chất: xảy ra khi thủ phạm truy cập bất hợp pháp vào tài khoản của một khách hàng hợp pháp. Gian lận chồng chất được coi là vấn đề quan trọng hơn và do đó nó thường là chủ đề cơ bản của các nỗ lực nghiên cứu". Kỹ thuật phát hiện sai lệch và phát hiện bất thường rất thường được áp dụng để phát hiện gian lận chồng chéo. Một số phương pháp được áp dụng gần đây bao gồm sử dụng kết hợp chữ ký của khách hàng, phân cụm động và phát hiện độ lệch; nhận dạng mẫu bằng cách sử dụng các công cụ trực quan để nhận dạng các mẫu bất thường. Một ứng dụng cụ thể đã được VNPT đưa ra giải pháp giúp chặn 200.000 cuộc gọi giả mạo mỗi tháng dựa trên dữ liệu về hành vi, đặc điểm cuộc gọi (các thông tin khai thác không phải thông tin riêng của người sử dụng), đầu code số điện thoại không đúng quy định với tiêu chuẩn quốc tế, đồng thời áp dụng các thuật toán dự đoán trên hạ tầng Big data, Máy học, Trí tuệ nhân tạo (AI) tới tất cả cuộc gọi nội mạng và ngoại mạng, VNPT xác định thuê bao phát tán cuộc gọi rác hoặc lừa đảo và ngăn chặn trong thời gian thực (VNPT, 2020).

Quản lý giám sát mạng: Để tăng trưởng và duy trì tỷ suất lợi nhuận trong ngành viễn thông đòi hỏi tăng hiệu quả mạng tối ưu và đảm bảo độ tin cậy của mạng ngày càng cao. Các công cụ phân tích trí tuệ doanh nghiệp

(Business Intelligence – BI) và KPDL đã được chứng minh là rất hiệu quả để so sánh một loạt các chỉ số hoạt động trong hệ thống mạng, tạo báo cáo thời gian thực để xác định các vấn đề cần chú ý ngay lập tức và tạo cảnh báo để thông báo tức thì về các tình huống khẩn cấp cần phản ứng nhanh. Quá trình KPDL cũng phân tích trình tự, dữ liệu chuỗi thời gian và phân loại chúng; tập trung vào việc ứng dụng các phương pháp khai phá dữ liệu để phát hiện những thay đổi về chất lượng dịch vụ dựa trên các phép đo mạng được tạo ra trong các hệ thống hoạt động. Với sự phát triển vượt bậc của công nghệ viễn thông hiện nay, mạng 5G đã được dần triển khai rộng rãi tại nhiều quốc gia trong đó có Việt Nam. Một trong những nhà sản xuất hệ thống di động hàng đầu thế giới Ericsson đã dẫn đầu một nghiên cứu để phân tích tính tiên tiến, kỳ vọng áp dụng AI, Máy học của các nhà khai thác mạng di động và các nhà cung cấp dịch vụ viễn thông toàn cầu. Nghiên cứu cho thấy các nhà khai thác sẽ chủ yếu sử dụng AI, Máy học như một công cụ để chuyển đổi sang hệ thống mạng 5G và đảm bảo tối ưu hóa đầu tư. Hơn nữa với 4G, 4.5G việc quản lý một số lượng lớn các thiết bị và lượng dữ liệu khổng lồ ngày càng trở nên phức tạp. Các nhà khai thác đang rất hy vọng ứng dụng AI và Máy học sẽ giúp giảm bớt sự phức tạp trong việc quản lý giám sát (Haidine & nnk., 2021).

Nhóm bài toán dự đoán hành vi thuê bao luôn là nhóm bài toán được quan tâm trong lĩnh vực viễn thông *(thuộc nhóm chủ đề "Tiếp thị, bán hàng và quan hệ khách hàng")*. Lợi ích đem lại từ kết quả KPDL trong lĩnh vực này hỗ trợ công tác tiếp thị, bán hàng và chăm sóc khách hàng nhằm:

- Tăng doanh thu dịch vụ.
- Giảm chi phí truyền thông.
- Giảm thuê bao rời mạng.
- Tiếp cận hợp lý và tránh làm gây phiền hà tới khách hàng.
- Chiếm lĩnh thị phần mảng dịch vụ viễn thông.

Để thực hiện được bài toán này việc tiếp cận các nguồn dữ liệu hữu ích của các hệ thống nghiệp vụ liên quan rất quan trọng. Việc phân tích và hiểu rõ dữ liệu thu thập là bước tiền đề để có được kết quả thành công cũng như ứng dụng mô hình đó vào trong công việc thực tế.

Trong nhóm bài toán này, chúng tôi lựa chọn một bài toán cụ thể để đưa vào thực nghiệm và đánh giá những giải thuật phù hợp cho nó là: *dự đoán hành vi hủy của thuê bao do chất lượng dịch vụ*. Với ý nghĩa thực tiễn giúp người quản trị, nhân viên quản lý dịch vụ, nhân viên quản lý địa bàn có thể ứng dụng, giám sát thường xuyên số liệu để có phương án kỹ thuật tối ưu kịp thời, chính sách thu hút hợp lý nhằm giữ chân khách hàng trước khi họ đưa quyết định hủy dịch vụ.

Song song với việc tìm hiểu bài toán là quá trình tiếp cận thu thập dữ liệu từ nhiều nguồn khác nhau để đảm bảo dữ liệu thu thập đúng và đủ cho bước thực nghiệm.

Đã có một vài nghiên cứu tại Việt Nam quan tâm tới việc dùng các mô hình học máy để KPDL trong tập khách hàng. Cụ thể, trong luận văn thạc sĩ tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội, tác giả Nguyễn Ngọc Tuân (Nguyễn Ngọc Tuân, 2016) đã sử dụng thuật toán cây quyết định C4.5 dự báo thuê bao rời mạng trong mạng di động. Tác giả Đoàn Văn Tâm (Đoàn Văn Tâm, 2019), trong luận văn thạc sĩ tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội đã xây dựng mô hình dự đoán khách hàng tiềm năng cho các gói cước trong mạng di động thông qua một số mô hình máy học: Decision trees, support vector machine, KNN, và kết hợp các mô hình phân lớp.

Trong bài báo này, chúng tôi lựa chọn và xử lý 6000 mẫu tập tin khác nhau được thu thập từ thực tế và tiến hành thực nghiệm trên 10 mô hình học máy cho bài toán phân lớp (classification problem): Logistic Regression (LR), Naïve Bayes (NB), K Neighbours Classifier (KNN), Decision Tree Classifier (DT), Random Forest Classifier (RF), Support Vector Classification (SVC), AdaBoost (AB), Bagging Classifier (BC), Extra Trees Classifier (ET), Gradient Boosting Classifier (GB). Tính tới thời điểm hiện tại, theo như hiểu biết của chúng tôi, bài báo này đã thực hiện tìm hiểu và áp dụng nhiều mô hình phân lớp nhất cho bài toán khai phá dữ liệu viễn thông.

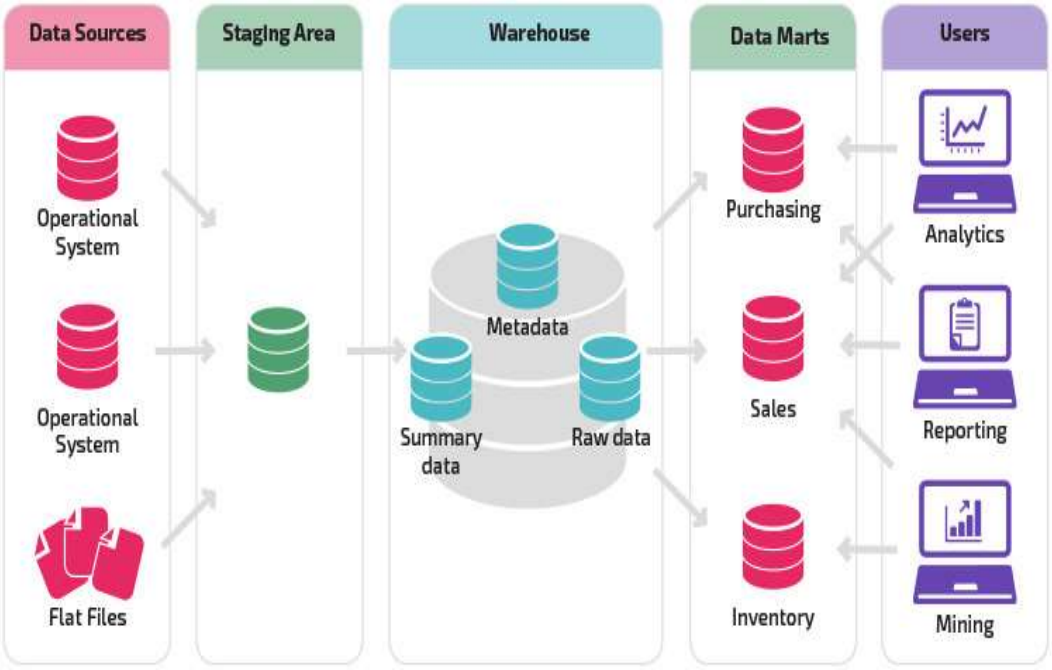
2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Thu thập dữ liệu

Dữ liệu thuê bao trong lĩnh vực viễn thông là một bài toàn dữ liệu đa dịch vụ và thường xuyên phải tái cấu trúc để phù hợp với hiện trạng thực tế triển khai. Về khối lượng dữ liệu dịch vụ khách hàng là rất lớn từ những nguồn dữ liệu chủ dịch vụ khác nhau, nó được tổng hợp và đồng bộ với nguồn dữ liệu tập trung, thống nhất dành cho việc xây dựng các hệ thống quản lý tập trung. Bởi vậy tham chiều dữ liệu thuê bao nó bao gồm nguồn dữ liệu tập trung là những dữ liệu dạng tổng hợp của thuê bao và dữ liệu phân tán là dữ liệu chi tiết về sử dụng dịch vụ của thuê bao đặt tại các hệ thống chủ dịch vụ.

Để thực hiện quy trình hóa việc cấp phát dữ liệu cho các bài toán, chúng tôi sử dụng mô hình data warehouse (DWH). Với DWH, việc kết nối với các nguồn dữ liệu tại Staging area, sau đó qua quá trình ETL (Extract – Transform – Load) dữ liệu sử dụng công cụ PDI (Pentaho Data Integration – mã nguồn mở của Hitachi) được lưu trữ thành 3 dạng metadata, summary data, raw data tại Warehouse, cuối cùng cấp phát dữ liệu cho các bài toán khai phá dữ liệu, báo cáo BI tại kho dữ liệu nhỏ gọn (data marts) (Fatima, 2019) dưới dạng dữ liệu chủ đề, các bước được mô tả như tại Hình 1.

Để phục vụ bài toán thực nghiệm, chúng tôi trích xuất dữ liệu 6000 mẫu dữ liệu (bao gồm 15 trường thông tin) theo dõi quá trình biến động trong 15 tháng, dữ liệu bao gồm bảng tổng hợp sự kiện (fact table) và một số bảng dữ liệu được lấy từ kho dữ liệu.



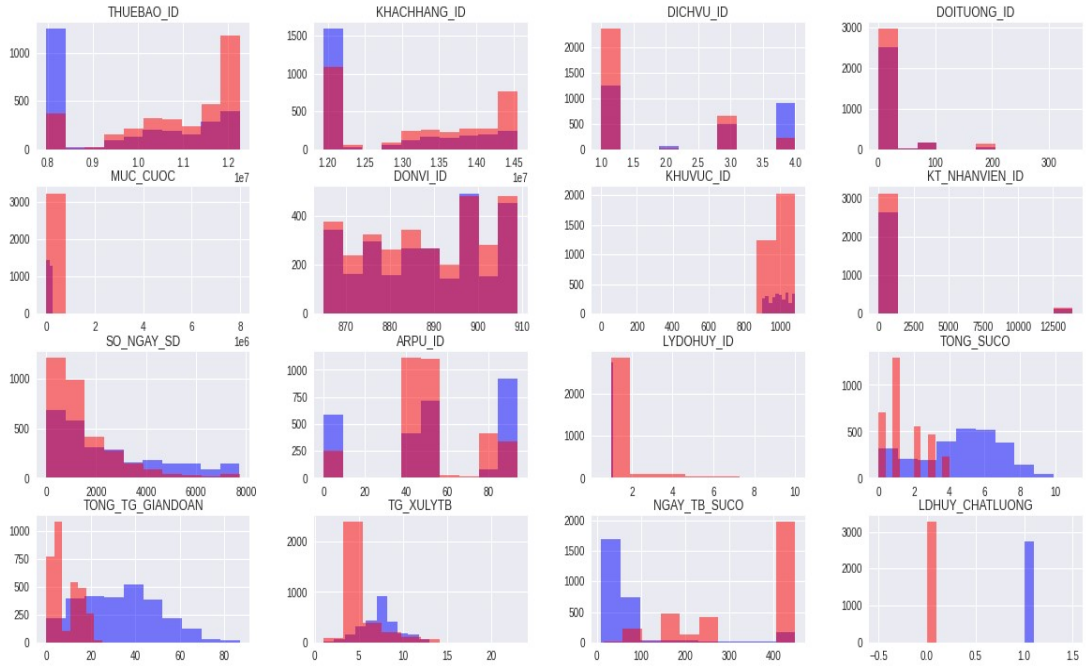
(Nguồn: panoply.io)

Hình 1: Mô hình Data wareHouse được áp dụng để quy trình hóa quá trình cấp phát dữ liệu cho các bài toán KPDL, BI

2.2. Trích chọn các đặc trưng quan trọng

Hình 2 thể hiện dữ liệu phân phối của từng trường với dữ liệu nhãn LDHUY_CHATLUONG cần phân loại cho bài thực nghiệm. Quan sát trên biểu đồ, phân phối có màu đỏ là phân phối của các đặc trưng của thuê bao đã bị hủy. Ngược lại, phân phối có màu xanh là phân phối của các đặc trưng của thuê bao chưa hủy. Trục hoành của biểu đồ bao gồm các giá trị khác nhau của các đặc trưng, và trục tung là tần suất xuất hiện của các giá trị ứng với trục hoành. Qua biểu đồ trên có thể thấy, phân phối các đặc trưng của hai nhãn hoàn toàn khác nhau. Ví dụ như đặc

trưng TG_XULYTB có thể thấy phần lớn các thuê bao bị hủy là do thời gian xử lý trung bình nằm trong khoảng 0 đến 5 ngày. Bên cạnh đó, những thuê bao không bị hủy thường có TG_XULYTB là từ 5 đến 10 ngày. Hoặc như đặc trưng NGAY_TB_SUCO, các thuê bao bị hủy có số ngày gặp sự cố trải dài từ 80 đến 400 ngày, và phần lớn thuê bao bị hủy đều gặp sự cố từ 400 ngày. Điều này cũng thật dễ hiểu vì càng gặp nhiều sự cố, càng có nhiều thuê bao bị hủy. Nói tóm lại, qua các phân phối trên, ta đánh giá được liệu mô hình máy học của ta có thể hoạt động tốt hay không.



Hình 2: Trực quan phân phối dữ liệu thu thập từng trường với dữ liệu nhãn LDHUY_CHATLUONG: đỏ-nhãn 0, xanh-nhãn 1

Tuy nhiên, 15 trường dữ liệu này ta không thể lấy và huấn luyện mô hình tất cả được. Một số đặc trưng có thể gây ra overfiting. Điển hình như đặc trưng THUEBAO_ID, với mỗi một ID như vậy ta có thể xác định được một thuê bao đã hủy dịch vụ hay không. Điều này là không tốt đối với các mô hình máy học. Đặc trưng LYDOHUY_ID cũng vậy, dựa vào phân phối trên có thể thấy nó mang nhiều giá trị khác nhau từ 0 (không hủy) đến 8 (hủy với các lý do khác nhau). Tuy nhiên, bài toán tập trung vào hành vi của khách hàng là có hủy do chất lượng dịch vụ hay không, từ đó có hướng đi phù hợp đối với các khách hàng tiếp tục sử dụng hoặc đã hủy các dịch vụ. Thay vì loại bỏ hoàn toàn đặc trưng này, ta xử lý lại nó khiến nó chỉ mang hai giá trị 0 (không hủy do chất lượng) và 1 (đã hủy do chất lượng dịch vụ), ta gọi đặc trưng đã được xử lý này là LDHUY_CHATLUONG.

Tiếp đến, thực hiện nạp và tiền xử lý dữ liệu thu thập được bằng một số kỹ thuật với ngôn ngữ python trên môi trường google colab. Sau đó thực hiện đánh giá thông qua hệ số tương quan Pearson (Pearson correlation coefficient) để xác định độ tương quan giữa các trường dữ liệu (Kent State University).

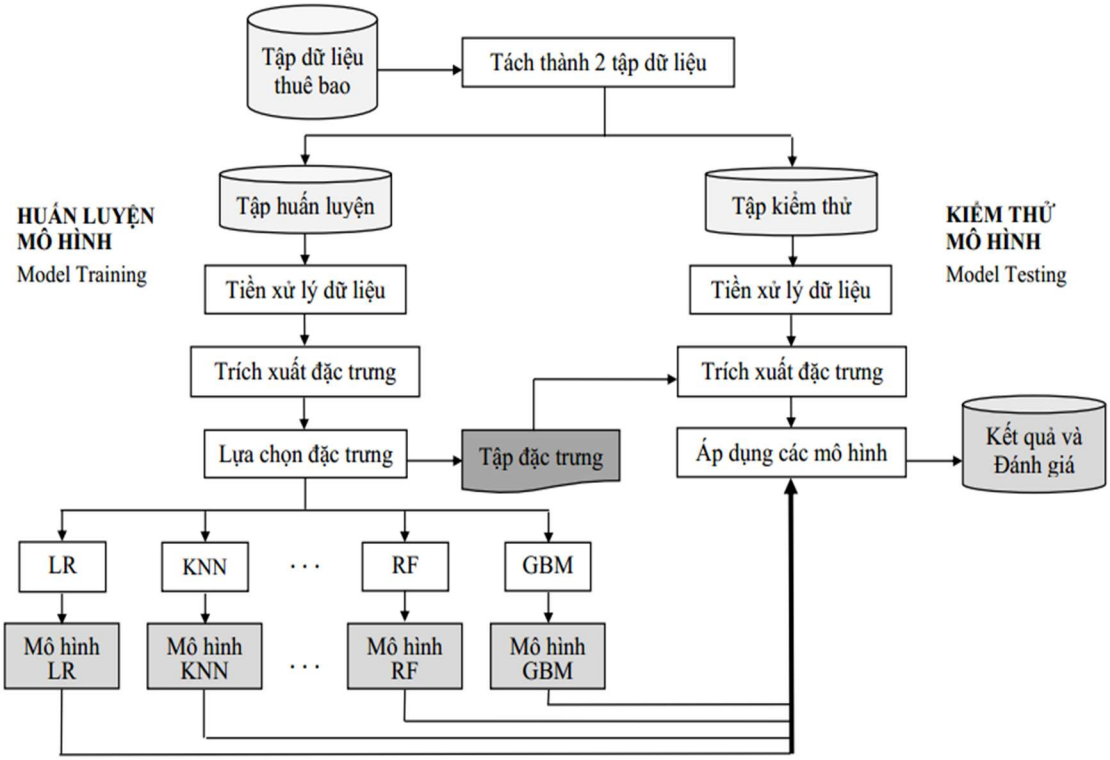
Sau khi đã có được các tính được hệ số tương quan giữa các trường dữ liệu với dữ liệu nhãn cần xác định là LDHUY_CHATLUONG. Đối với 2 đặc trưng là DONVI_ID và KT_NHANVIEN_ID mang giá trị tương quan rất thấp, lần lượt là 0.008511 và -0.002382. Qua đó trích chọn ra một số trường dữ liệu có mối tương quan với kết quả để lấy làm đặc trưng quan trọng cho bài thực nghiệm, bao gồm 11 đặc trưng như Bảng 1.

Bảng 1. Các đặc trưng quan trọng trích chọn được

TT	Đặc trưng	Mô tả
1	KHACHHANG_ID	ID khách hàng, một khách hàng có thể dùng nhiều thuê bao
2	DICHVU_ID	ID dịch vụ
3	DOITUONG_ID	ID của đối tượng khách hàng
4	MUC_CUOC	Mức cước của thuê bao
5	KHUVUC_ID	Địa bàn quản lý
6	SO_NGAY_SD	Số ngày sử dụng đến lúc hủy
7	ARPU_ID	ID mức doanh thu trung bình của khách hàng
8	TONG_SUCO	Số lượng sự cố, hỏng hóc khách hàng báo trong thời gian lấy mẫu
9	TONG_TG_GIANDOAN	Tổng thời gian gián đoạn dịch vụ trong thời gian lấy mẫu
10	TG_XULYTB	Thời gian trung bình sửa chữa của thuê bao trong thời gian lấy mẫu
11	NGAY_TB_SUCO	Số ngày trung bình phát sinh sự cố từ thời diểm sử dụng dịch vụ

2.3. Mô hình thực nghiệm

Các bước thực hiện của mô hình thực nghiệm trên Hình 3 được mô tả như sau: Với tập dữ liệu thu thập được sẽ được tách làm 2 tập huấn luyện và tập kiểm thử với tỉ lệ 80/20. Dữ liệu 2 tập này lần lượt qua các bước tiền xử lý dữ liệu, trích xuất đặc trưng và quan trọng là trích chọn được đặc trưng quan trọng (mục 2.2 đã nêu). Rồi lần lượt áp dụng 10 giải thuật để tìm ra 10 mô hình máy học tương ứng trên dữ liệu tập huấn luyện. Sau đó áp dụng 10 mô hình đó với dữ liệu kiểm thử để cho ra kết quả phân loại của từng mô hình, cuối cùng sử dụng ma trận đánh giá confusion matrix (sẽ đề cập tại mục 3.1) để đánh giá kết quả của từng mô hình thông qua các độ đo.



Hình 3: Mô hình các bước thực hiện bài toán thực nghiệm

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Ma trận đánh giá (confusion matrix)

Chúng tôi sử dụng phương pháp đánh giá thông qua các độ đo của ma trận đánh giá (confusion matrix) (Stehman, 1997) để đánh giá kết quả thu được của các mô hình. Bảng 2 dưới đây là các chỉ số của confusion matrix trong quá trình so sánh giữa dữ liệu thực tế với dữ liệu dự đoán.

Bảng 2. Ma trận đánh giá (confusion matrix)

	Positive Dự đoán	Nevative Dự đoán
Positive Thực tế	True Positive (TP)	False Negative (FN)
Negative Thực tế	False Positive (FP)	True Negative (TN)

Từ những chỉ số trên, confusion matrix đã đưa ra các độ đo như: Accuracy, Precision, Recall, F-score (F1) để đánh giá sự hiệu quả của một mô hình máy học. Cụ thể ý nghĩa của một số độ đo như sau:

Accuracy (độ chính xác): Số dữ liệu do mô hình dự đoán đúng / Tổng số dữ liệu mô hình dự đoán *(Bao gồm cả positive và negative)*.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

− Precision: Số dữ liệu do mô hình dự đoán đúng positive / Tổng số dữ liệu mô hình dự đoán positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

− Recall (Độ hồi tưởng): Số dữ liệu mô hình dự đoán đúng positive / Tổng số dữ liệu thực tế positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

− F-score (F1): Là độ đo hài hòa giữa độ chính xác và độ hồi tưởng của mô hình.

$$\text{F1} = \frac{2 \text{ x Recall x Precision}}{\text{Recall} + \text{Precision}}$$

Trong bài toán thực nghiệm chúng tôi sử dụng độ hồi tưởng (recall) để đánh giá kết quả của các mô hình, ở phần kết quả và đánh giá chúng tôi sẽ giải thích lý do lựa chọn độ đo này trong việc đánh giá mô hình.

3.2. Thực nghiệm

Sau khi đã trích chọn đặc trưng ở mục 2.1, tiếp đến tiến hành tách dữ liệu ra thành hai tập dữ liệu: huấn luyện/tập kiểm tra là 80/20 (huấn luyện là 4800 khách hàng và kiểm tra là 1200 khách hàng).

Để tăng độ chính xác và sự hội tụ cho mô hình máy học, chúng tôi thực hiện chuẩn hóa dữ liệu (Normalization) để đưa dữ liệu của các thuộc tính về đoạn [0, 1]. Các mô hình máy học trong thực nghiệm đều sử dụng thư viện Sklearn.

Chúng tôi cài đặt bài toán thực nghiệm với 10 thuật toán phân lớp trên thư viện sklearn (Sklearn , 2021), bao gồm: Logistic Regression (LR), Naïve Bayes (NB), K Neighbours Classifier (KNN), Decision Tree Classifier (DT), Random Forest Classifier (RF), Support Vector Classification (SVC), AdaBoost (AB), Bagging Classifier (BC), Extra Trees Classifier (ET), Gradient Boosting Classifier (GB). Chúng tôi cũng để các tham số mặc định. Tại thời điểm chúng tôi thực nghiệm, sklearn đang có phiên bản *scikit-learn0.24.2*.

3.3. Kết quả và đánh giá

Với kết quả thực nghiệm có được trên các mô hình máy học trên bộ dữ liệu thu thập, thông qua confusion matrix có được các độ đo của các mô hình và chúng tôi đặt giả định positive là hủy do chất lượng dịch vụ tức nhãn dự đoán LDHUY_CHATLUONG = 1, lúc đó các độ đo cần tổng hợp được ở Bảng 3 phía dưới. Quan sát trực quan bảng trên, chúng ta nhận thấy kết quả các độ đo của cả 10 mô hình giải thuật đều cao, điều này cho thấy mô hình máy học là cách tiếp cận đầy hứa hẹn cho bài toán dự đoán hành vi thuê bao trong lĩnh vực viễn thông.

Bảng 3 cho chúng ta thấy rằng mô hình Gradient Boosting cho kết quả tốt nhất trên cả ba độ đo Accuracy, Recall, và F-score. Thực nghiệm này cũng phù hợp về mặt lý thuyết

khi mô hình Gradient Boosting là một trong những mô hình mạnh mẽ nhất trong các mô hình học máy.

Đặc biệt, với ý nghĩa bài toán thực nghiệm này các đơn vị viễn thông sẽ quan tâm tìm được mô hình với kết quả dự đoán hủy với lý do *"Chất lượng dịch vụ"* với xác suất bỏ sót ít nhất (*Thà báo nhầm còn hơn bỏ sót*). Do vậy, các đơn vị viễn thông sẽ quan tâm độ đo Recall hơn so với các độ đo khác. Chính vì vậy, mô hình Gradient Boosting có độ hồi tưởng cao nhất là 0.986 sẽ là mô hình lý tưởng để tiếp cận dự đoán hành vi của khách hàng cho bài toán viễn thông được đề cập trong bài báo này.

Bảng 3. Các số đo kết quả của 10 mô hình máy học (Số liệu bôi đậm chỉ kết quả tốt nhất của độ đo)

Mô hình	*Accuracy*	*Precision*	*Recall*	*F1-score*
LR	0.925	0.918	0.911	0.915
NB	0.893	0.884	0.874	0.879
KNN	0.944	0.914	0.956	0.935
DT	0.948	**0.954**	0.930	0.942
RF	0.966	0.947	0.975	0.960
SVC	0.958	0.933	0.970	0.952
AB	0.960	0.937	0.970	0.954
BA	0.961	0.943	0.967	0.955
ET	0.966	0.952	0.969	0.961
GB	**0.968**	0.939	**0.986**	**0.962**

4. KẾT LUẬN

Hiện này vẫn còn ít những nghiên cứu và áp dụng các mô hình máy học để khai phá dữ liệu trong bài toán viễn thông. Qua quá trình tiếp cận, tìm hiểu dữ liệu chúng tôi đã đưa ra phương án quy trình hóa việc cấp phát dữ liệu doanh nghiệp viễn thông cho các bài toán thông minh doanh nghiệp (BI) và khai phá dữ liệu.Tiếp đó, chúng tôi cũng tiến hành thực nghiệm trên dữ liệu thực tế thu thập được để dự đoán hành vi thuê bao khách hàng thông qua 10 giải thuật phân lớp: Logistic Regression, Naïve Bayes, K Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Classification, AdaBoost, Bagging Classifier, Extra Trees Classifier, Gradient Boosting Classifier. Tính tới thời điểm hiện tại, theo như những hiểu biết của chúng tôi, bài báo này đã thực hiện tìm hiểu và áp dụng nhiều mô hình phân lớp nhất cho bài toán khai phá dữ liệu viễn thông.

Chúng tôi rút ra được hai kết luận sau:

i) Dùng mô hình máy học vào khai phá dữ liệu, cụ thể trong bài toán dự đoán hành vi của khách hàng – muốn hủy dịch vụ hay không – là phù hợp;

ii) Mô hình Gradient Boosting có độ hồi tưởng (recall) là rất cao – 0.986.

Với những gì đạt được, chúng tôi sẽ tiếp tục nghiên cứu về bài toán này. Cụ thể, chúng tôi sẽ tiếp tục thu thập dữ liệu và đa dạng hóa dữ liệu. Tiếp nữa, chúng tôi còn dự định phân tích đặc trưng nào sẽ mang lại nhiều đóng góp quan trọng trong bài toán này. Từ đó thuyết phục tổ chức sử dụng những phát hiện của mô hình máy học để đưa ra các điều chỉnh phù hợp với khách hàng, và chiến lược phát triển kinh doanh.

TÀI LIỆU THAM KHẢO

Bloice, M. D., & Holzinger, A. (2016). A Tutorial on Machine Learning and Data Science Tools with Python. *Lecture Notes in Computer Science*, 435–480. DOI:10.1007/978-3-319-50478-0_22.

Đoàn Văn Tâm. (2019). *Xây dựng mô hình dự đoán khách hàng tiềm năng cho các gói cước trong mạng di động*. Luận văn Thạc sĩ, Đại học Công nghệ, Đại học Quốc gia Hà Nội.

Fatima, N. (2019). Data Warehouse Architecture: Types, Components, & Concepts. Truy cập từ https://www.astera.com/type/blog/data-warehouse-architecture/ abgerufen.

Garg, R. (2018). 7 Types of Classification Algorithms. Truy cập từ https://analyticsindiamag.com/7-types-classification-algorithms/ abgerufen

Haidine, A., Salmam, F. Z., Aqqal, A., & Dahbi, A. (2021). Artificial Intelligence and Machine Learning in 5G and beyond: A Survey and Perspectives. In *Moving Broadband Mobile Communications Forward - Intelligent Technologies for 5G and Beyond.* DOI:10.5772/intechopen.98517

Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection . *Knowledge-Based Systems, 21*(7), 721–726. doi:10.1016/j.knosys.2008.03.026

Kabakchieva, D. (2009). Business Intelligence Applications and Data Mining Methods in Telecommunications: A Literature Review. Bulgaria. Truy cập từ:https://core.ac.uk/download/pdf/71542208.pdf abgerufen

LibGuides: SPSS Tutorials: Pearson Correlation. (n.d.). (20. July 2021). Uni. Von Kent State University. Truy cập từ https://libguides.library.kent.edu/spss/pearsoncorr. abgerufen

Maimon, O., & Rokach, L. (2005). Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.

Nguyễn Ngọc Tuân. (2016). *Áp dụng kỹ thuật KPDL dự báo thuê bao rời mạng trong mạng di động.* Luận văn Thạc sĩ, Đại học Công nghệ, ĐHQGHN.

Singh, H. (2018). Understanding Gradient Boosting Machines - Towards Data Science. *Medium*. Truy cập 04/11/2018 từ https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab. abgerufen

Sklearn. (2021). Supervised learning. Truy cập từ: https://scikit-learn.org/stable/supervised_learning.html abgerufen

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment, 62* (1), 77–89.

Supervised learning. (n.d.). Scikit-Learn. Truy cập 20/07/2021 từ https://scikit-learn.org/stable/supervised_learning.html abgerufen

VNPT ứng dụng giải pháp AI chặn 200.000 cuộc gọi giả mạo mỗi tháng. (2020). *Công Nghệ và Đời Sống*. Truy cập từ http://congnghevadoisong.vn/vnpt-ung-dung-giai-phap-ai-chan-200000-cuoc-goi-gia-mao-moi-thang-d34218.html abgerufen

Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data, 7*(1). doi:10.1186/s40537-020-00290-0

Weiss, G. (2009). Data Mining in Telecommunications.

THÔNG TIN TÁC GIẢ

TS. Nguyễn Văn Hậu

- Nghiên cứu sinh về biểu diễn và lập luận tri thức tại khoa Khoa học máy tính, Trường Đại học Kĩ thuật Dresden (Technische Universitäte Dresden – TUD), CHLB Đức.

- Trưởng khoa Công nghệ Thông tin, Trường Đại học Sư phạm Kỹ thuật Hưng Yên. Giám đốc Trung tâm AI.

- Lĩnh vực nghiên cứu: Boolean satisfiability problems (SAT), Constraint Satisfaction Problems (CSPs); Data Mining, Machine Learning, Artificial Intelligence.