CASE STUDY

DỰ BÁO KHÁCH HÀNG THUÊ BAO RỜI MẠNG DỊCH VỤ FIBER

Võ Đức Vinh[1], Trần Văn Lăng[2,*]

[1] VNPT Đồng Nai
[2] Trường Đại học Ngoại ngữ -Tin học TP.HCM
vdvinh.dni@vnpt.vn, langtv huflit.edu.vn

TÓM TẮT— Bài báo này trình bày áp dụng công cụ trong lĩnh vực học máy và thuật toán cây quyết định để xây dựng mô hình phân tích dữ liệu dự báo sớm những thuê bao rời mạng. Mô hình sử dụng nguồn dữ liệu lịch sử các thuộc tính là nguyên nhân gây ra sự rời mạng của các thuê bao tại VNPT Đồng Nai. Kết quả dự báo khách hàng có khả năng rời mạng cao với tỉ lệ chính xác dự báo của mô hình so với số liệu thực tế rất cao.

Từ khóa— Học máy, cây quyết định, khách hàng thuê bao, mạng viễn thông.

I. GIỚI THIỆU

Thuê bao rời mạng luôn là vấn đề "đau đầu" của các nhà mạng trong nước cũng như trên thế giới bởi lẽ khách hàng (thuê bao) chính là người mang lại doanh thu và duy trì hoạt động của các nhà mạng. Để duy trì và phát triển hoạt động kinh doanh của mình, các nhà mạng phải tìm mọi cách để phát triển thuê bao mới nhưng đồng thời cũng phải tìm cách để duy trì hoạt động của các thuê bao hiện hữu. Tổng chi phí để phát triển một thuê bao mới cao hơn nhiều so với việc duy trì một thuê bao hiện hữu.

Để phát triển 1 thuê bao mới các khoản chi phí phải tốn:

- Chi phí nhân công: 50.000/thuê bao;
- Dây thuê bao: Đơn giá * chiều dài mét dây bình quân theo định mức = 900 * 150 = 135.000 VNĐ;
- Modem: 900.000 VNĐ.
- Tổng chi phí cho 1 thuê bao phát triển mới khoảng: 1.085.000 VNĐ.

Trong khi đó việc thực hiện chính sách khuyến mãi giảm giá cước đối với khách hàng sử dụng lâu năm (VD: giảm giá cước 3 kỳ hoá đơn tháng sau liền kề…) nhằm mục đích giữ chân khách hàng hiện hữu sẽ khỏi phải mất khoản hao hụt chi phí như trên, vả lại tạo niềm tin đối với khách hàng.

Bên cạnh đó, doanh thu từ các thuê bao hiện hữu (đặc biệt là các thuê bao lâu năm) cao hơn nhiều so với doanh thu của các thuê bao mới (theo thống kê số liệu tại VNPT Đồng Nai, thuê bao lâu năm có doanh thu trung bình cao hơn so thuê bao mới: doanh thu bình quân của thuê bao trước ngày 31/12/2020 là 169.952 VNĐ, thuê bao phát triển mới trong năm 2021 là 123.485 VNĐ). Chính vì lý do trên, các nhà mạng trên thế giới không ngừng tìm kiếm các giải pháp và nghiên cứu phát triển các ứng dụng để xác định, dự đoán sớm thuê bao có khả năng rời mạng để có biện pháp kịp thời tác động nhằm duy trì thuê bao đó hoạt động.

Ngày nay, thị trường viễn thông trên toàn thế giới đang phải đối mặt với mất doanh thu nghiêm trọng do cạnh tranh gay gắt và mất khách hàng tiềm năng. Để giữ lợi thế cạnh tranh và có được càng nhiều khách hàng càng tốt, hầu hết các nhà khai thác đầu tư một khoản chi khổng lồ để mở rộng hoạt động kinh doanh của họ. Do đó, việc giữ chân khách hàng tiềm năng (khách hàng hiện hữu) trở nên quan trọng đối với các nhà khai thác để có thể thu lại số tiền đã đầu tư và đạt được lợi nhuận trong một khoảng thời gian ngắn nhất.

Việc khách hàng ngừng sử dụng dịch vụ của công ty trong một khoảng thời gian nhất định và chuyển sang nhà mạng khác được định nghĩa là khách hàng rời mạng [1].

Các công ty thì luôn muốn có thêm càng nhiều khách hàng càng tốt. Mặc dù vậy, qua thời gian, tỷ lệ khách hàng mới/ khách hàng rời mạng có xu hướng tiến tới bằng 1. Vì vậy, tác động của khách hàng rời mạng ngày càng trở nên mạnh mẽ và cần được quan tâm hơn.

Việc rời mạng thường gắn liền với vòng đời của ngành, khi ngành đang trong giai đoạn phát triển, việc bán hàng tăng trưởng bùng nổ, số khách hàng mới vượt xa số khách hàng rời mạng, nhưng khi ở giai đoạn bão hòa, các công ty sẽ tập trung vào việc giảm tỉ lệ rời mạng.

Thời điểm khách hàng rời mạng sẽ cho biết khách hàng gắn bó với công ty trong bao lâu, giá trị vòng đời của khách hàng (CLV) đối với công ty. CLV được tính bằng tổng doanh thu mà Công ty thu được từ khách hàng trong suốt vòng đời của khách hàng trừ đi tổng chi phí thu hút khách hàng, bán hàng, dịch vụ khách hàng (quy ra tiền).

Các nghiên cứu trước đây đã đưa ra khái niệm khách hàng rời mạng từ nhiều quan điểm khác nhau. Theo Olafsson, Li, và Wu [3], có 2 loại rời mạng khác nhau. Loại thứ nhất là rời mạng chủ động (nghĩa là khách hàng

* Coressponding Author

chủ động chọn dừng sử dụng dịch vụ). Loại thứ hai là rời mạng bị động (nghĩa là khi những khách hàng không còn là khách hàng tốt nữa và công ty lựa chọn dừng mối quan hệ này).

Burez và Van den Poel [2] đã chia rời mạng chủ động thành 2 nhóm: Rời mạng do vấn đề thương mại và rời mạng do vấn đề tài chính. Rời mạng do vấn đề thương mại là trường hợp khách hàng không gia hạn hợp đồng có thời hạn cố định của họ khi hợp đồng hết hạn. Rời mạng do vấn đề tài chính là trường hợp khách hàng ngừng thanh toán trong quá trình thực hiện hợp đồng mà họ đang bị ràng buộc.

Ngày nay, khách hàng rời mạng đã trở thành vấn đề quan tâm chính của các công ty trong tất cả các lĩnh vực và các công ty buộc phải hành động để xử lý vấn đề này.

Xem xét tỷ lệ rời mạng của các ngành khác nhau, có thể nhận thấy ngành viễn thông là một trong những ngành có tỉ lệ rời mạng cao nhất với tỉ lệ rời mạng trung bình hàng năm từ 20% đến 40%. Khách hàng rời mạng trong lĩnh vực viễn thông được hiểu là khách hàng chuyển từ nhà cung cấp này sang nhà cung cấp khác.

Có 2 cách tiếp cận cơ bản đối với việc quản lý rời mạng. Cách tiếp cận thứ nhất là tiếp cận không có mục tiêu dựa vào các sản phẩm nổi trội và truyền thông rộng rãi để tăng lòng trung thành và duy trì khách hàng. Cách tiếp cận thứ hai là tiếp cận có mục tiêu dựa vào việc xác định những khách hàng có khả năng rời mạng, sau đó cung cấp cho họ những giá trị khuyến khích trực tiếp (khuyến mại) hoặc tạo ra các gói dịch vụ phù hợp cho khách hàng để giữ họ ở lại.

Cách tiếp cận có mục tiêu gồm 2 loại: bị động và chủ động. Với cách tiếp cận bị động, công ty chờ cho đến khi khách hàng liên hệ với công ty để hủy dịch vụ, công ty sau đó mới đưa ra cho khách hàng những chính sách khuyến khích, ví dụ khuyến mại giảm giá, để giữ khách hàng ở lại. Với cách tiếp cận chủ động, công ty cố gắng xác định những khách hàng có khả năng rời mạng trong một thời gian ngắn tiếp theo. Sau đó, công ty sẽ thực hiện những chương trình hoặc chính sách đặc biệt để giữ cho khách hàng không rời mạng. Cách tiếp cận chủ động có những lợi ích là chi phí khuyến khích thấp (bởi vì phần khuyến khích đó không cần thiết phải cao như tại thời điểm khách hàng đã quyết định sẽ rời mạng rồi) và bởi vì khách hàng không được chuẩn bị sẵn để thương lượng có được mức khuyến khích tốt hơn với lý do rời mạng. Tuy nhiên, cách tiếp cận này sẽ gây lãng phí nếu việc dự đoán rời mạng là không chính xác, bởi vì sau đó công ty sẽ phải lãng phí một lượng lớn chi phí để khuyến khích những khách hàng thực tế vẫn ở lại với mạng mình.

Để giải quyết vấn đề này, rất nhiều nỗ lực đã thực hiện để có được cái nhìn chính xác hơn về rời mạng. Nhìn chung, các nghiên cứu trong lĩnh vực này đều tập trung về một trong những mục đích chính sau: tìm ra các nhân tố ảnh hưởng đến khách hàng rời mạng, hoặc xây dựng mô hình cho việc dự đoán khách hàng rời mạng.

Hiện nay, trong lĩnh vực phát triển dịch vụ FiberVnn trong nước có ba nhà mạng viễn thông lớn đang đầu tư trong lĩnh vực này là Tập đoàn bưu chính viễn thông Việt Nam "VNPT", Công ty viễn thông FPT, Công ty viễn thông Viettel. Theo số liệu của Cục viễn thông công bố thị phần đến cuối năm 2021, thị phần internet cáp quang của VNPT đứng thứ 2 chiếm 32,31% thị phần, đứng thứ 1 là Viettel chiếm 48,96%, FPT chiếm 18,5% thị phần còn lại là của các doanh nghiệp khác. Do đó việc canh tranh và thu hút khách hàng trong việc phát triển thuê bao cũng như giữ chân khách hàng của các nhà mạng cạnh tranh ngày càng trở nên quyết liệt hơn.

Tuy nhiên, hiện nay bài toán dự báo khách hàng rời mạng đã được các Viễn thông tỉnh, thành phố khác đã xây dựng và đưa vào sử dụng chủ yếu phân tích và dự báo trên dịch vụ di động chưa áp dụng cho dịch vụ Fiber.

Hiện tại, Viễn thông Đồng Nai thực hiện phân tích số liệu để lọc danh sách các khách hàng sử dụng dịch vụ Fiber có khả năng rời mạng cao chủ yếu dựa vào các chỉ tiêu báo cáo thống kê từ các hệ thống điều hành sản xuất kinh doanh, hệ thống BI cũ dẫn đến việc dự báo số liệu không được nhanh chóng, chính xác và mất rất nhiều thời gian.

Xuất phát từ những khó khăn và yêu cầu đặt ra đối với đơn vị mình, nhóm đề tài nghiên cứu xây dựng hệ thống cơ sở dữ liệu khách hàng, tập hợp lịch sử các thuộc tính là nguyên nhân ảnh hưởng đến sự rời mạng của khách hàng (Ví dụ: độ hài lòng của khách hàng trong công tác lắp đặt và sửa chữa; việc khách hàng thực hiện thanh toán đúng hạn; độ ổn định phục vụ dịch vụ của nhà mạng…) và áp dụng các kỹ thuật máy học vào việc phân tích dữ liệu khách hàng đã rời mạng từ đó dự báo thuê bao đang sử dụng có khả năng rời mạng cao.

II. PHƯƠNG PHÁP THỰC HIỆN

Trong bài báo này, phương pháp thực hiện bài toán dự báo thuê bao rời mạng bằng thuật toán cây quyết định tăng cường hai lớp và nguồn số liệu thực tế tại VNPT Đồng Nai.

Tập dữ liệu mẫu dùng để huấn luyện là tập dữ liệu lịch sử các thuộc tính là nguyên nhân gây ảnh hưởng đến việc rời mạng của khách hàng được lưu trữ và trích xuất đến tháng 12/2020 của 238.700 thuê bao đang sử dụng hoặc đã rời mạng tại VNPT Đồng Nai, bao gồm 14 cột trong đó 13 cột đầu bao gồm các thuộc tính ảnh hưởng đến sự rời mạng của thuê bao, cột 14 là cột Churn (thanhly), cột này là cột gắn nhãn của tập dữ liệu, cột để nhận biết là thuê bao có ý định rời mạng hay không.

Các thuộc tính sử dụng trong tập dữ liệu huấn luyện:

- Khu vực (khuvuc_id): địa bàn (ấp, xã, khu phố, xã phường…) huyện quản lý thuê bao.
- Đối tượng (phanloaikh_id): đối tượng phân loại khách hàng (cá nhân, doanh nghiệp, hành chính sự nghiệp, trường học…).
- Số lần báo hỏng (solan_bh): Số lần thuê bao báo hỏng do sự cố (đứt cáp, không tín hiệu, mạng chập chờn …).
- Số lần gọi kiểm (solan_gk): Số lần bộ phận chăm sóc khách hàng thực hiện gọi kiểm để khảo sát dịch vụ đường truyền trong việc lắp đặt và sửa chữa.
- Số lần gọi kiểm hài lòng (solan_gk_hl): Số lần khách hàng trả lời hài lòng khi được gọi kiểm.
- Số lần gọi kiểm không hài long (solan_gk_khl): Số lần khách hàng trả lời hài không lòng khi được gọi kiểm.
- Số lần tạm ngưng (solan_td): Số lần khách hàng xin tạm ngưng hoặc bị tạm ngưng sử dụng dịch vụ (do yêu cầu hoặc nợ cước …).
- Số tháng sử dụng (sothang_sd): Tuổi đời sử dụng của khách hàng.
- Giá cước (gia_cuoc): Giá gói cước khách hàng đăng ký sử dụng trọn gói trong tháng.
- Không phát sinh lưu lượng (kpsll): Số ngày không phát sinh lưu lượng của thuê bao tháng trước liền kề.
- Số lần gia hạn đặt cọc (solan_gh_datcoc): Số lần thuê bao thực hiện gia hạn đặt cọc trả trước khi hết tiền đặt cọc.
- Số tháng sử dụng khi hết đặt cọc (sothang_sd_hetdc): Số tháng khách hàng sử dụng tiếp sau khi hết tiền đặt cọc (chuyển qua hình thức trả sau).
- Điểm tín nhiệm (diemtinnhiem): Số điểm tính nhiệm đánh giá khách hàng (Các tiêu chí đánh giá: thời gian sử dụng dịch vụ, giá gói cước dịch vụ, thanh toán tiền đúng hạn…).
- Thanh lý (thanhly): Trạng thái thuê bao còn sử dụng hoặc thanh lý.

Mô hình dự báo sẽ được thực hiện bằng hai phương pháp:

A. PHƯƠNG PHÁP 1

Phương pháp này sẽ thực hiện trên bộ dự liệu thô ban đầu chưa qua bước tiền xử lý và chuẩn hoá dữ liệu, bao gồm 13 thuộc tính của tập danh sách 238.700 khách hàng được lưu trữ dưới dạng file .CSV (mau_train.csv), tỉ lệ số khách hàng dán nhãn (0:1) như Bảng 1.

Bảng 1. *Bảng tỉ lệ số khách hàng dán nhãn (0:1) chưa thực hiện tiền xử lý*

Tổng số record	Tổng số gán nhãn (0)	Tỉ lệ	Tổng số gán nhãn (1)	Tỉ lệ
238700	171.561	71.87%	67139	28.13%

Bảng 2. *Bảng tập dữ liệu chưa chuẩn hoá*

khuvuc_id	phanloaikh_id	solan_bh	solan_gk	solan_gk_hl	solan_gk_khl	solan_td	sothang_sd	muccuoctb_id	kpsll	solan_gh_datcoc	sothang_sd_hetdc	diemtinnhiem	thanhly
432	15	0	1	0	0	0	7	130817	0	0	0	50	0
528	15	1	2	2	0	0	7	130817	0	0	0	50	0
391	15	0	1	1	0	0	7	120000	0	0	0	50	0
366	15	0	1	1	0	0	7	120000	0	0	0	50	0
542	15	0	1	0	0	0	7	171818	0	0	0	50	0
525	15	0	1	0	0	0	7	130817	0	0	0	50	0
371	8	0	1	1	0	0	7	190909	0	0	0	54	0
516	15	0	1	0	0	1	7	0	0	0	0	50	0
512	16	0	1	1	0	0	7	199091	0	0	0	50	0
391	8	0	1	1	0	0	7	327273	0	0	0	0	0
418	15	0	1	1	0	0	7	0	0	0	0	46	0
389	15	0	1	1	0	0	7	120000	0	0	0	50	0
606	15	0	1	0	0	0	7	130817	0	0	0	50	0
613	15	1	2	1	1	0	7	157181	0	0	0	50	0
502	15	1	2	2	0	0	7	171818	0	0	0	50	0
353	15	0	1	1	0	0	7	0	0	0	0	46	0
406	15	0	0	0	0	0	7	171818	0	0	0	46	0
601	15	0	0	0	0	0	10	130817	0	0	0	55	0
601	15	0	0	0	0	0	10	130817	0	0	0	55	0
601	15	0	1	1	0	0	10	130817	0	0	0	55	0
601	15	0	0	0	0	0	10	171818	0	0	0	55	0
563	15	0	1	0	0	0	7	171818	0	0	0	50	0
447	15	0	1	0	1	0	7	0	0	0	0	46	0
526	15	0	1	0	1	0	7	157272	0	0	0	50	0
420	15	0	1	1	0	0	7	171818	0	0	0	46	0
466	15	0	1	1	0	0	7	120000	0	0	0	46	0
406	8	0	1	1	0	0	7	645455	0	0	0	74	0
351	15	1	2	2	0	0	7	199090	0	0	0	50	0
542	15	0	1	1	0	1	7	171818	0	0	0	46	0

Để thực hiện huấn luyện mô hình chúng tôi sử dụng phần mềm Microsoft Azure (Machine Learning) với các bước thao tác như sau:

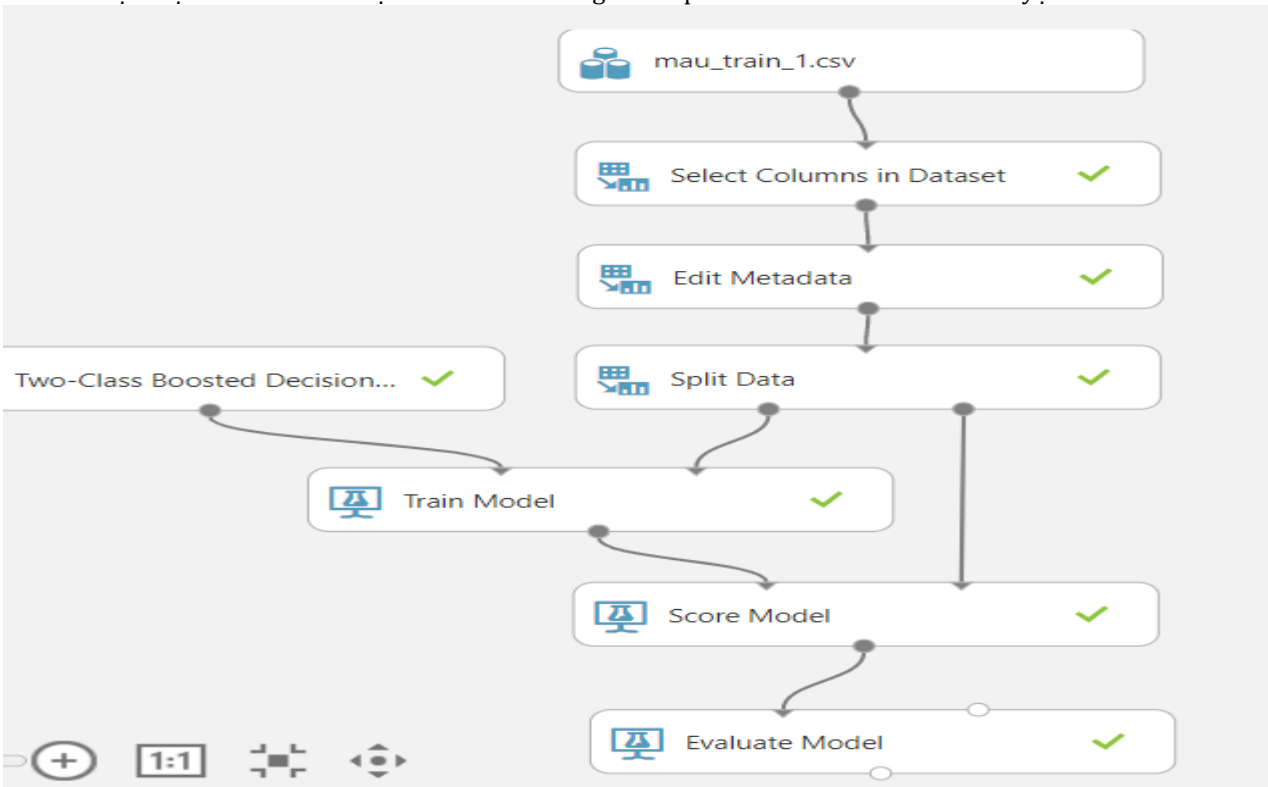
- Bước 1: Thực hiện upload file dữ liệu mẫu huấn luyện (mau_train_1.csv)

  Chọn tab "Datasets": là nơi dùng để lưu trữ và quản lý dữ liệu. Sau khi hoàn tất thực hiện upload file ta sẽ thấy xuất hiện file như Hình 1:

Hình 1. *Vùng chứa Datasets*

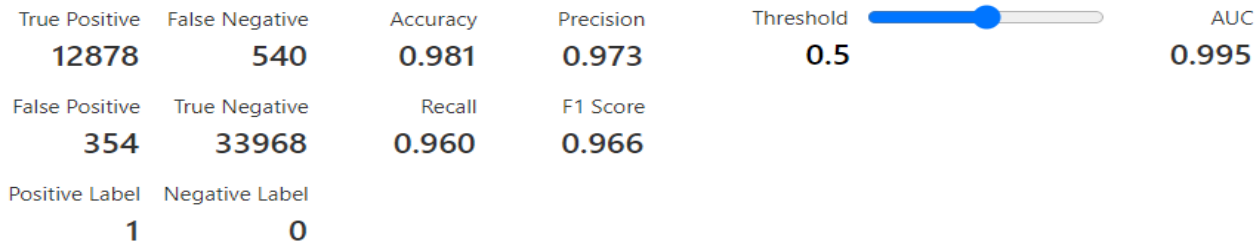
- Bước 2: Chọn Tab "Experiments": Có chức năng để thực hiện xây dựng mô hình. Tại bước này ta thực hiện kéo thả trình tự các công cụ đã dựng sẵn và cài đặt cấu hình cho từng thành phần:
  - Datasets "mau_train_2": Quản lý tập dữ liệu mẫu.
  - Select Column in Dataset: Có tính năng để chọn các thuộc tính để huấn luyện.
  - Edit Metadata module: Có tính năng xử lý các dữ liệu bị lỗi.
  - Split Data: Có tính năng chia tập dữ liệu mẫu thành hai tập dữ liệu dùng để huấn luyện và dữ liệu kiểm tra mẫu (mặc định được cài đặt 80:20).
  - Two-Class Boosted Decision Tree module: Thuật toán cây quyết định tăng cường hai lớp là phương pháp học tổng hợp, trong đó cây thứ hai sửa lỗi cho cây đầu tiên. Dự đoán dựa trên toàn bộ nhóm cây cùng nhau đưa ra dự đoán. Cây quyết định tăng cường hai lớp là thuật toán dễ dàng và đạt được hiệu quả tốt trong bài toán dự báo.
  - Train Model: Thực hiện huấn luyện mô hình.
  - Score Model: Đánh giá mô hình dự báo.
  - Evaluate Model: Kết suất giá trị mô hình.

Sau khi thực hiện kết nối và cài đặt cấu hình cho từng thành phần ta sẽ có mô hình huấn luyện như Hình 2.



Hình 2. *Mô hình huấn luyện đã cài đặt cấu hình và kết nối*

- Bước 3: Thực hiện lưu trữ phần cài đặt, chạy huấn luyện mô hình và chúng ta được kết quả các chỉ tiêu dự báo của mô hình như Hình 3.



Hình 3. *Kết quả các chỉ số của mô hình huấn luyện*

B. PHƯƠNG PHÁP 2:

Phương pháp 2 sẽ thực hiện chuẩn hoá lại bộ dữ liệu huấn luyện mẫu ban đầu trước khi huấn luyện mô hình. Chúng ta thực hiện các giai đoạn sau:

1. Giai đoạn 1: Tiền xử lý dữ liệu

Theo như bộ dữ liệu thu thập mẫu ban đầu có tổng số thuê bao là 238.700 với tổng thuê bao dán nhãn "0" là 171.561: 71.87% và tổng thuê bao dán nhãn "1" là 67.139: 28.13% tạo ra sự mất cân bằng dữ liệu dẫn đến mô hình sẽ dự đoán nghiên về trường hợp thuê bao sẽ sử dụng nhiều hơn. Để xử lý việc mất cân bằng dữ liệu như trên chúng ta sẽ phải cập nhật và dán nhãn "1" lại cho các thuê bao được xem là có nguy cơ rời mạng cao bằng cách dựa vào số liệu thống kê cho các tình huống như sau:

- Số lần tạm ngưng >=3 lần (thuê bao có nợ cước >=3 lần nguy cơ rời mạng cao).
- Điểm tính nhiệm <=75 và giá gói cước sử dụng >300.000 (khách hàng sử dụng ngắn hạn nguy cơ rời mạng cao).
- Thuê bao có số lần báo hỏng >3 và số tháng sử dụng >24 (khách hàng không hài lòng nhà mạng nguy cơ rời mạng cao).
- Thuê bao có giá cước = 0 (lỗi dữ liệu).

Sau khi cập nhật và xử lý lại, tập dữ liệu mẫu mới có tổng số thuê bao là 238.700 với tổng thuê bao dán nhãn "0" là 143.539 : 60% và tổng thuê bao dán nhãn "1" là 95.161 : 40%.

Sử dụng kỹ thuật feature scaling để chuẩn hóa dữ liệu cho giá trị của các feature thuộc [-1, 1] (Độ dốc gradient hội tụ nhanh hơn là khi không chuẩn hóa dữ liệu).

Ta sử dụng công thức của feature scaling vào hàm tiền xử lý dữ liệu:

$$\text{X\_new} = \frac{Xi - Xmean}{Standard\ Deviation}$$

Sau khi thực hiện feature scaling ta nhận được bảng kết quả Bảng 3.

Bảng 3. *Bảng dữ liệu sau khi tiền xử lý*

khuvuc_id	phanloaikh_id	solan_bh	solan_gk	solan_gk_hl	solan_gk_khl	solan_td	sothang_sd	gia_cuoc	kpsll	solan_gh_datcoc	sothang_sd_hetdc	diemtinnhiem	thanhly
-0.38124707	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.096194075	-0.084738899	0	0	-0.906474812	0
0.801261058	0.087897799	0.470507817	1.113326222	1.601602608	-0.389268579	-0.364741604	-1.272488711	-0.096194075	-0.084738899	0	0	-0.906474812	0
-0.886276583	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.118667619	-0.084738899	0	0	-0.906474812	0
-1.194221407	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.118667619	-0.084738899	0	0	-0.906474812	0
0.97371016	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.01100985	-0.084738899	0	0	-0.906474812	0
0.764307679	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.096194075	-0.084738899	0	0	-0.906474812	0
-1.132632443	-3.966736663	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	0.028653866	-0.084738899	0	0	-0.583809832	0
0.653447542	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	1.17918264	-1.272488711	-0.36798122	-0.084738899	0	0	-0.906474812	0
0.60417637	0.667131293	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	0.045652899	-0.084738899	0	0	-0.906474812	0
-0.886276583	-3.966736663	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	0.311965533	-0.084738899	0	0	-4.939787069	0
-0.553696172	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.36798122	-0.084738899	0	0	-1.229139793	0
-0.910912169	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.118667619	-0.084738899	0	0	-0.906474812	0
1.762048912	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.096194075	-0.084738899	0	0	-0.906474812	0
1.848273463	0.087897799	0.470507817	1.113326222	0.361710079	1.898733363	-0.364741604	-1.272488711	-0.041419877	-0.084738899	0	0	-0.906474812	0
0.48099844	0.087897799	0.470507817	1.113326222	1.601602608	-0.389268579	-0.364741604	-1.272488711	-0.01100985	-0.084738899	0	0	-0.906474812	0
-1.354352716	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.36798122	-0.084738899	0	0	-1.229139793	0
-0.701509688	0.087897799	-0.422304636	-1.011293445	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.01100985	-0.084738899	0	0	-1.229139793	0
1.700459947	0.087897799	-0.422304636	-1.011293445	-0.878182449	-0.389268579	-0.364741604	-1.056601152	-0.096194075	-0.084738899	0	0	-0.503143586	0
1.700459947	0.087897799	-0.422304636	-1.011293445	-0.878182449	-0.389268579	-0.364741604	-1.056601152	-0.096194075	-0.084738899	0	0	-0.503143586	0
1.700459947	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.056601152	-0.096194075	-0.084738899	0	0	-0.503143586	0
1.700459947	0.087897799	-0.422304636	-1.011293445	-0.878182449	-0.389268579	-0.364741604	-1.056601152	-0.01100985	-0.084738899	0	0	-0.503143586	0
1.232383813	0.087897799	-0.422304636	0.051016388	-0.878182449	-0.389268579	-0.364741604	-1.272488711	-0.01100985	-0.084738899	0	0	-0.906474812	0
-0.196480175	0.087897799	-0.422304636	0.051016388	-0.878182449	1.898733363	-0.364741604	-1.272488711	-0.36798122	-0.084738899	0	0	-1.229139793	0
0.776625472	0.087897799	-0.422304636	0.051016388	-0.878182449	1.898733363	-0.364741604	-1.272488711	-0.041230814	-0.084738899	0	0	-0.906474812	0
-0.529060586	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.01100985	-0.084738899	0	0	-1.229139793	0
0.037557892	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	-0.118667619	-0.084738899	0	0	-1.229139793	0
-0.701509688	-3.966736663	-0.422304636	0.051016388	0.361710079	-0.389268579	-0.364741604	-1.272488711	0.973024702	-0.084738899	0	0	1.029515071	0
-1.378988302	0.087897799	0.470507817	1.113326222	1.601602608	-0.389268579	-0.364741604	-1.272488711	0.045650821	-0.084738899	0	0	-0.906474812	0
0.97371016	0.087897799	-0.422304636	0.051016388	0.361710079	-0.389268579	1.17918264	-1.272488711	-0.01100985	-0.084738899	0	0	-1.229139793	0

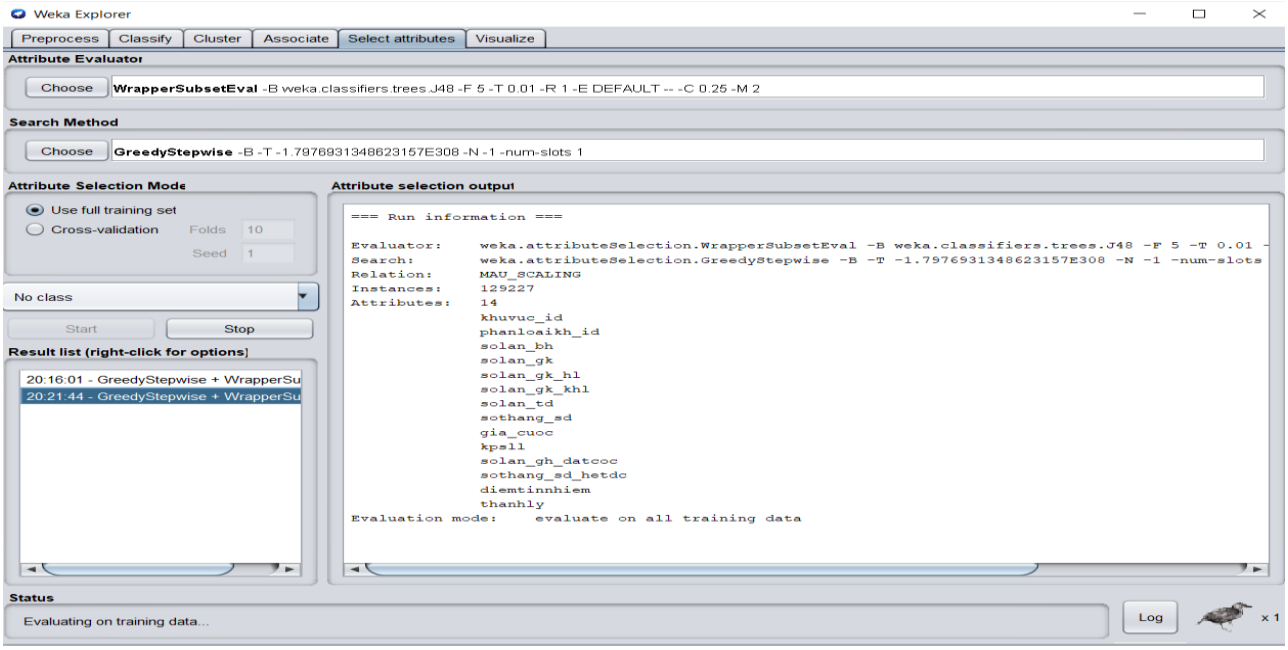
2. Giai đoạn 2: Rút trích dữ liệu

Có rất nhiều trường thông tin dữ liệu được chiết xuất và tổng hợp ở các giai đoạn trước, nhưng ở giai đoạn này chỉ rút trích một số trường dữ liệu nhất định phục vụ cho việc giải bài toán. Phần mềm Weka được sử dụng để trích chọn các thuộc tính, nhằm mục đích loại bỏ các thuộc tính dư thừa, thu gọn tập dữ liệu mẫu, tạo tiền đề quan trọng cho cải tiến hiệu năng, tốc độ xử lý và độ chính xác của tập dữ liệu đầu ra cho mô hình cây quyết định.

Trong Weka, đề tài sẽ sử dụng chức năng Attribute Elevator để thực hiện lựa chọn thuộc tính thông qua hai phương pháp được cấu hình và thực thi trên Weka.

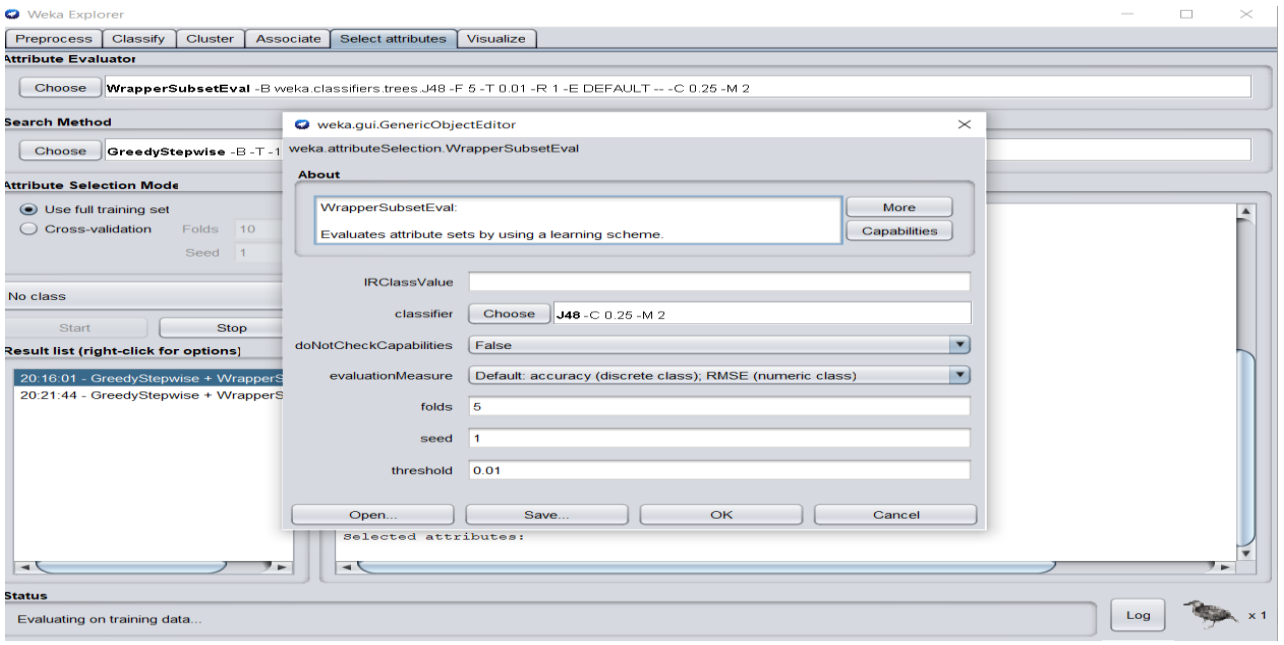
Các bước thực hiện:

- Bước 1: chọn tab "Select attributes": Đây là công cụ có tính năng thực hiện lựa chọn các thuộc tính có ảnh hưởng đến tập dữ liệu huấn luyện mẫu, tránh dư thừa thuộc tính tập dữ liệu mẫu giúp cho mô hình dự báo có độ chính xác tốt hơn như Hình 4.



Hình 4. *Lựa chọn thuộc tính trên WeKa*

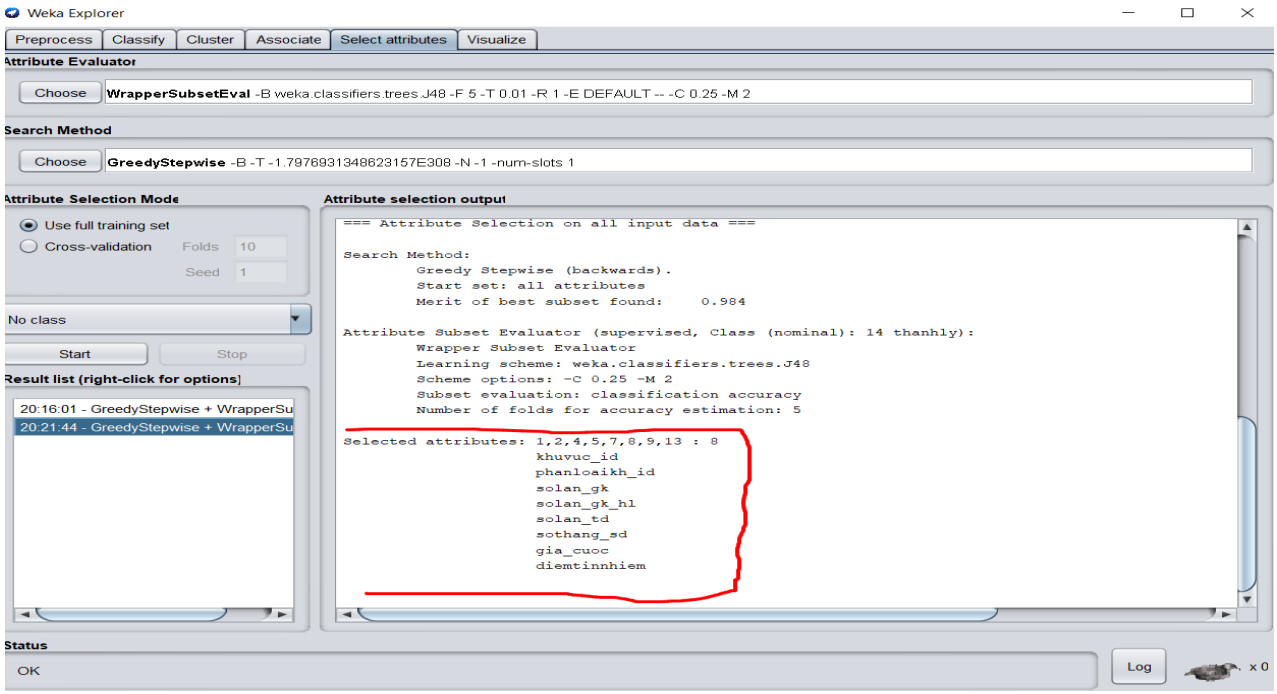
- Bước 2: Tuỳ thuộc vào giải thuật dự định chọn để sử dụng cho quá trình phân loại sau này, mà ta sẽ chọn giải thuật tương ứng để trích chọn thuộc tính. Giải thuật "J48" trong Weka được chọn, bởi đây là phương pháp phân lớp bằng mô hình cây thay thế giải thuật C4.5. Việc tìm ra được thuộc tính nào tốt nhất cho quá trình phân loại sau này được thể hiện như Hình 5.



Hình 5. *Chọn giải thuật trên Weka*

Sử dụng phương pháp trích chọn thuộc tính như trên ta thu được tập dữ liệu huấn luyện mẫu mới với những thuộc tính được chọn lọc có tác động đến bộ dữ liệu mẫu cho bài toán dự báo. Chúng ta biết rằng, kết quả của việc chọn thuộc tính phụ thuộc rất lớn vào tập huấn luyện (training dataset). Nếu sử dụng một dataset khác sẽ có thể thu thập được tập thuộc tính khác có khi các kết quả rất khác nhau.

- Bước 3: Sau khi chọn phương pháp + giả thuật và thực thi chương trình, kết quả sẽ cho ra các thuộc tính quan trọng như Hình 6:



Hình 6. *Kết quả chọn ra các thuộc tính quan trọng*

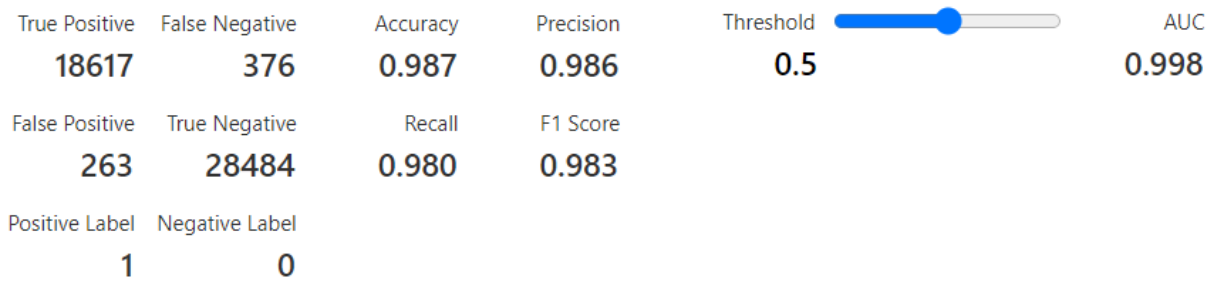
Sau khi thực hiện chuẩn hoá và trích lọc các thuộc tính quan trọng, bảng dữ liệu tập huấn luyện mới được lưu trữ với tên file có định dạng .CSV (mau_train_2.CSV) có 9 thuộc tính (khuvuc_id, phanloaikh_id, solan_gk, solan_gk_hl, solan_td, sothang_sd, gia_cuoc, diemtinnhiem, thanhly) như Bảng 4:

Bảng 4. *Bảng dữ liệu mẫu huấn luyện mới sau khi chuẩn hoá và trích lọc thuộc tính*

khuvuc_id	phanloaikh_id	solan_gk	solan_gk_hl	solan_td	sothang_sd	gia_cuoc	diemtinnhiem	thanhly
0.579540784	0.087897799	0.051016388	-0.878182449	-0.364741604	-1.200526192	-0.01100985	-0.906474812	1
-0.024031073	0.667131293	-1.011293445	-0.878182449	-0.364741604	1.03031192	-0.01100985	0.706850091	0
1.170794848	0.667131293	-1.011293445	-0.878182449	-0.364741604	1.318161999	-0.36798122	0.706850091	0
1.577282017	-3.966736663	-1.011293445	-0.878182449	1.17918264	0.022836644	0.338407318	0.868182581	1
-1.539119611	0.087897799	0.051016388	0.361710079	-0.364741604	-0.624826034	-0.118667619	0.142186375	0
-0.800052032	0.087897799	1.113326222	-0.878182449	-0.364741604	-0.480900994	-0.118667619	0.142186375	0
0.60417637	0.087897799	0.051016388	0.361710079	-0.364741604	-1.272488711	-0.118667619	-0.906474812	1
-0.639920723	0.087897799	0.051016388	0.361710079	-0.364741604	-1.200526192	-0.041419877	-0.664476077	1
1.762048912	0.087897799	0.051016388	0.361710079	-0.364741604	-1.056601152	0.045652899	-0.341811096	0
-0.541378379	0.087897799	0.051016388	0.361710079	-0.364741604	0.310686723	-0.014786951	0.303518865	0
-0.295022519	0.087897799	0.051016388	0.361710079	-0.364741604	0.598536802	-0.0091213	0.5455176	0
0.85053223	0.087897799	0.051016388	-0.878182449	-0.364741604	-1.272488711	0.045652899	-1.551804773	1
-0.812369825	0.087897799	0.051016388	0.361710079	-0.364741604	-1.344451231	-0.36798122	-1.713137263	1
0.554905198	0.087897799	-1.011293445	-0.878182449	-0.364741604	0.238724203	-0.014786951	0.5455176	0
-0.713827481	0.087897799	-1.011293445	-0.878182449	-0.364741604	0.454611762	0.026765316	0.5455176	0
-0.184162382	0.087897799	0.051016388	-0.878182449	-0.364741604	-0.768751073	-0.156064659	0.06152013	0
-0.787734239	0.087897799	0.051016388	0.361710079	-0.364741604	0.166761683	-0.014786951	0.5455176	0
0.653447542	0.667131293	0.051016388	0.361710079	-0.364741604	0.166761683	0.026765316	0.22285262	1
-0.824687618	0.087897799	0.051016388	0.361710079	-0.364741604	-0.768751073	-0.096194075	0.142186375	0
-1.144950236	0.087897799	1.113326222	0.361710079	-0.364741604	0.814424361	-0.118667619	0.706850091	0
-0.38124707	0.087897799	0.051016388	-0.878182449	-0.364741604	0.598536802	-0.046896466	0.5455176	0
-0.01171328	0.087897799	-1.011293445	-0.878182449	-0.364741604	1.606012078	-0.056339218	1.190847562	0

3. Giai đoạn 3: Xây dựng và huấn luyện mô hình

Thực hiện các bước 1, 2, 3 như trên giải pháp 1. Ta được kết quả các chỉ số dự báo của mô hình như Hình 7.



Hình 7. *Kết quả chỉ số mô hình dự báo của phương pháp 2*

4. Giai đoạn 4: Chuẩn bị số liệu để kiểm tra tỉ lệ dự báo của mô hình với số liệu thực tế

Dữ liệu dùng để kiểm tra mô hình dự báo được trích xuất từ dữ liệu khách hàng đang sử dụng dịch vụ Fiber tại VNPT Đồng Nai tính đến tháng 1/2021 với tổng số thuê bao là 165.000 thuê bao.

Dữ liệu được trích xuất theo mẫu huấn luyện với 8 thuộc tính được thống kê bắt đầu từ năm 1/2015 đến 1/2021.

III. KẾT QUẢ THỬ NGHIỆM

Công cụ sử dụng:

- Phần mềm WEKA (Phiên bản 3.9.5) được sử dụng để thu giảm tập dữ liệu huấn luyện.
- Microsoft Azure (Machine Learning): Đăng ký tài khoản Free-WorkSpace (trả phí theo cấu hình CPU: 2.45$/giờ).

A. KẾT QUẢ CỦA HAI PHƯƠNG PHÁP

Để đánh giá mô hình dự báo có độ chính xác tốt hay không, chúng cần xem xét các chỉ số sau:

Do yếu tố rời mạng là quan trọng trong dự báo, nên Positive đây là khả năng rời mạng.

- Precision: để đo độ chính xác (tỷ lệ phần trăm) của việc dự đoán đúng trong tất cả các dự đoán khách hàng rời mạng (bao gồm dự đoán đúng – true positive và dự đoán sai false positive).

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

+ TP (True Positive): Số thuê bao rời mạng được mô hình dự đoán đúng.

+ FP (False Positive): Số thuê bao rời mạng mô hình dự đoán sai.

- Recall: nhằm xác định tỷ lệ phần trăm của việc dự đoán đúng trong trong tất cả các trường hợp thực tế khách hàng đã rời mạng (bao gồm cả dự đoán đúng - true positive và dự đoán sai - false negative).

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

+ TP (True Positive): Số thuê bao rời mạng được mô hình dự đoán đúng.

+ FN (False Negative): Số thuê bao sử dụng được mô hình dự đoán sai.

- Accuracy: tỉ lệ phần trăm được mô hình dự đoán đúng trong trong tất cả các trường hợp khách hàng đang sử dụng và đã rời mạng của tập dữ liệu kiểm thử. Accuracy càng cao thì mô hình dự đoán càng chính xác.

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

+ TP (True Positive): Số thuê bao rời mạng được mô hình dự đoán đúng.

+ FP (False Positive): Số thuê bao rời mạng mô hình dự đoán sai.

+ FN (False Negative): Số thuê bao sử dụng được mô hình dự đoán sai.

+ TN (False Negative): Số thuê bao sử dụng được mô hình dự đoán đúng.

- AUC (Area Under The Curve): tỉ lệ phần trăm của mô hình dự đoán đúng trong tất cả trường hợp khách hàng đã rời mạng và khách hàng đang sử dụng.

Đánh giá kết quả sau khi thực hiện hai giải pháp 1 và 2 cho ta thấy kết quả của giải pháp 2 tốt hơn giải pháp 1, cũng đồng nghĩa việc chuẩn hoá và tiền xử lý dữ liệu mẫu ban đầu rất quan trọng. Giúp việc dự báo mô hình có kết quả tốt hơn. Bảng đánh giá các chỉ số của hai phương pháp như Bảng 5.

Bảng 5. *Bảng đánh giá các chỉ tiêu của hai phương pháp*

Phương pháp	Accuracy	Precision	Recall	AUC
Phương pháp 1: sử dụng tập dữ liệu thô chưa thực hiện cân bằng dữ liệu tập khách hàng gán nhãn (0:1) bị lệch rất lớn theo tỉ lệ (71.87% : 21.13%) và chưa trích lọc các thuộc tính quan trọng (13 thuộc tính)	**0.982**	**0.974**	**0.961**	**0.995**
Phương pháp 2: tập dữ liệu đã được thực hiện cân bằng dữ liệu tập khách hàng gán nhãn (0:1) theo tỉ lệ (60% : 40%) và đã trích lọc thuộc tính quan trọng (8 thuộc tính)	**0.987**	**0.986**	**0.980**	**0.998**
Tỉ lệ chênh lệch kết quả dự báo của phương pháp 1 và 2	**0.005**	**0.012**	**0.019**	**0.003**

B. SO SÁNH VỚI DỮ LIỆU THỰC TẾ

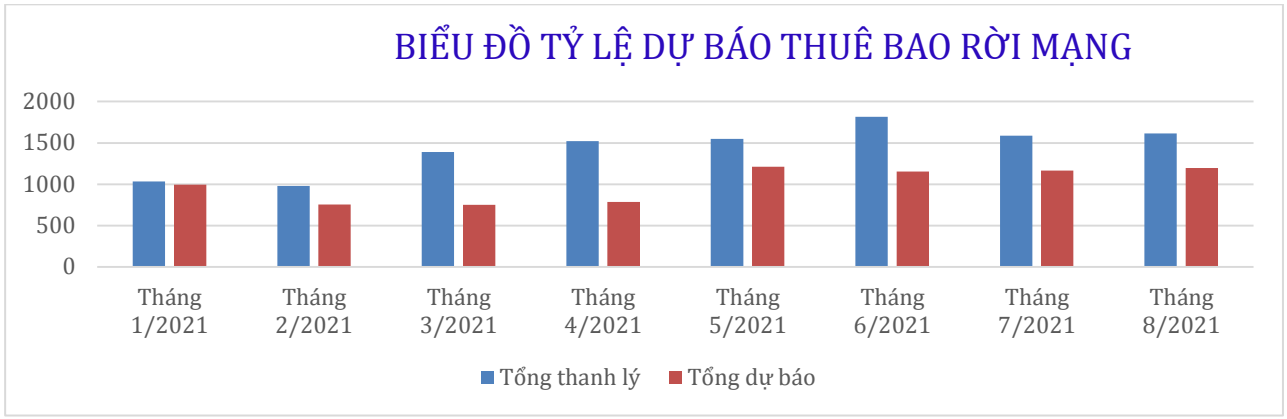
Kết quả dự báo của mô hình được đối soát với số liệu thuê bao thanh lý thực tế tại Viễn thông Đồng Nai qua các tháng như Bảng 6, biểu đồ dự báo thuê bao rời mạng như Hình 8.

Bảng 6. *Bảng thống kê số liệu dự báo so với số liệu thực tế từng tháng*

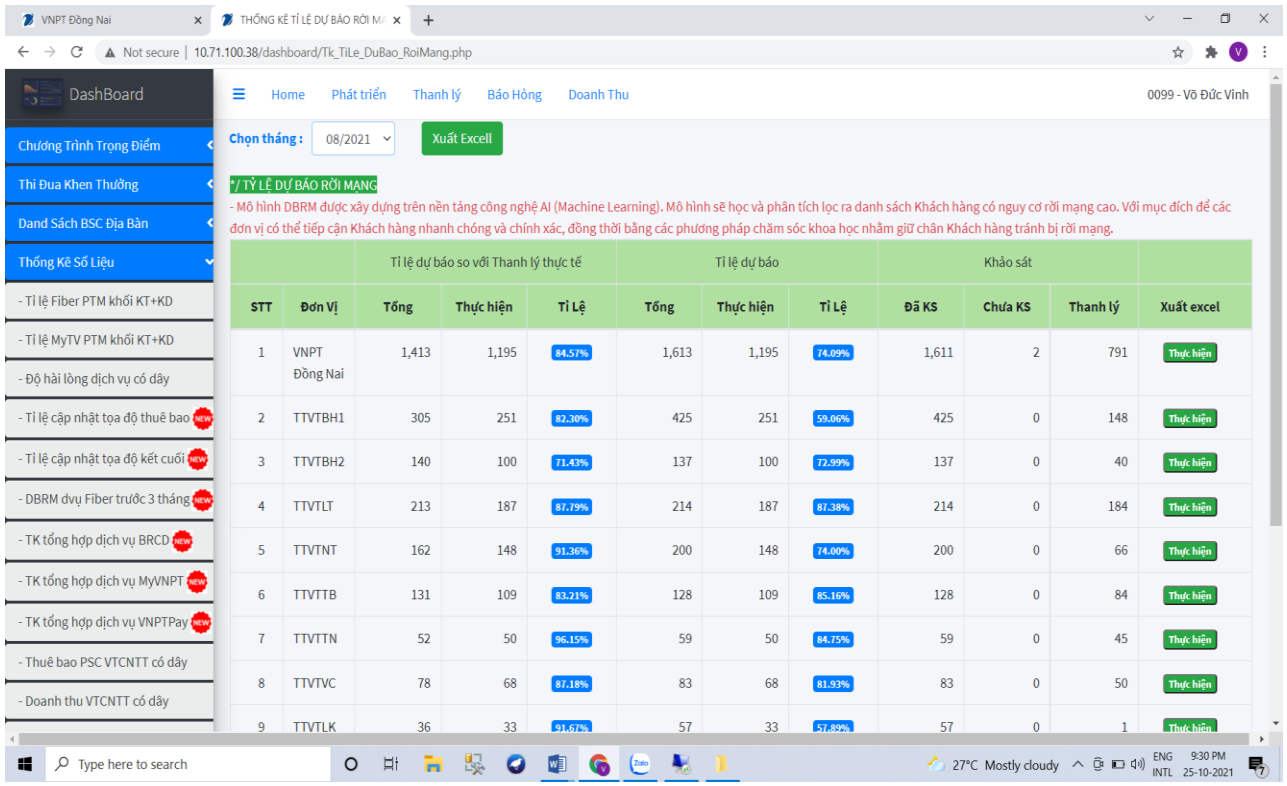
Tháng	Mô hình dự đoán	Thanh lý thực tế	Dự đoán đúng	Tỉ lệ
1/2021	1034	1252	994	96.13%
2/2021	981	991	757	77.17%
3/2021	1392	1014	751	53.95%
4/2021	1520	1191	788	51.84%
5/2021	1550	1590	1213	78.26%
6/2021	1817	1513	1152	63.40%
7/2021	1587	1537	1165	73.41%
8/2021	1613	1413	1195	74.09%

Tại sao các đại lượng accuracy, precision, recall đều cao (>95%), nhưng tỷ lệ dự đoán đúng so với thực tế trong Bảng 6 lại thấp do các nguyên nhân sau:

- Với mục đích giảm thiểu thuê bao rời và muốn giữ chân khách hàng nên có những trường hợp bất quy tắc trong việc thực hiện thanh lý thuê bao mặc dù thuê bao đủ điều kiện thanh lý.
- Khi chuẩn hoá dữ liệu với ý đồ sẽ dự báo khách hàng có khả năng rời mạng càng nhiều càng tốt để tránh dự báo thiếu sót.



Hình 8. *Biểu đồ tỷ lệ dự báo so với số liệu thực tế*



Hình 9. *Trang web thống kê tỷ lệ dự báo so với số liệu thực tế*

Đánh giá kết quả so với thực tế:

- Tháng 1/2021: có tỉ lệ dự báo chính xác rất cao do trước đây chưa có triển khai mô hình dự báo thuê bao rời mạng nên việc chăm sóc khách hàng gặp khó khăn không biết khách hàng nào có nguy cơ rời mạng cao để thực hiện chăm sóc và giữ chân khách hàng.
- Tháng 2, 3, 4, 5, 6, 7, 8/2021: Do có chủ trương giảm thuê bao rời mạng và đã được dự báo sớm khách hàng có nguy cơ rời mạng cao, nhờ đó bộ phận chăm sóc khách hàng đã có chiến lược tổ chức chăm sóc khách hàng tốt hơn. Nên tỉ lệ dự báo chính xác của mô hình thấp.

IV. KẾT LUẬN

Với cơ sở dữ liệu lưu trữ các thông tin về khách hàng như báo hỏng, gọi kiểm, thanh lý, danh bạ khách hàng, danh sách thuê bao không phát sinh lưu lượng, danh sách tập khách hàng trả trước sắp hết hạn; để từ đó xây dựng tập dữ liệu các thuộc tính là nguyên nhân ảnh hưởng đến nguy cơ rời mạng của khách hàng làm tập dữ liệu huấn luyện cho mô hình dự báo rời mạng.

Đồng thời tập hợp và chuẩn hóa bộ dữ liệu mẫu huấn luyện, do đặc thù của tập dữ liệu liên tục, dữ liệu quá lệch nhãn 0 nhiều hơn nhãn 1 nên cần xử lý bằng các phương pháp chuẩn hóa và chọn lựa đặc trưng để có tỉ lệ tập train, test tốt, tránh tình trạng học quá dư thừa (over fitting learning) ở mô hình cây quyết định.

Kết quả cho thấy mô hình dự báo với thuật toán Cây quyết định tăng cường hai lớp với có độc chính xác cao (99.8%).

V. LỜI CẢM ƠN

Cảm ơn Trung tâm Công nghệ thông tin của VNPT Đồng Nai đã hỗ trợ và cung cấp số liệu để thử nghiệm và kiểm tra tính đúng đắn của mô hình.

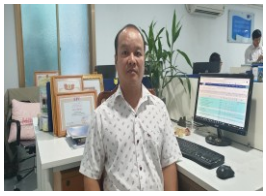
VI. TÀI LIỆU THAM KHẢO

[1] M. Chandar, A. Laha & P. Krishna (2006), "Modeling churn behavior of bank customers using predictive data mining techniques", National conference on soft computing techniques for engineering applications.
[2] J. Burez, & D. Van den Poel (2009), "Handling class imbalance in customer churn prediction", Expert System with Applications, 36, 4626-4636.
[3] S. Olafsson, X. Li, & S. Wu (2008), "Operations research and data mining", European Journal of Operational Research, 187, 2592-1448.
[4] Hội đồng Thành viên Tập Đoàn Bưu chính Viễn thông Việt Nam, "Bộ tài liệu Văn hóa VNPT", QĐ số 65/QĐ-VNPT ngày 5/5/2014.

FORECAST SUBSCRIBER CUSTOMERS CANCEL FIBER SERVICES OF TELECOMMUNICATION NETWORK

Vo Duc Vinh, Tran Van Lang

ABSTRACT— This paper presents the use of methods in the field of machine learning and decision tree algorithms to build a data analysis model that predicts early subscribers leaving the telecommunication network. The model of using historical data sources is the cause of the cancellation of subscribers at VNPT Dong Nai. The results predict those customers who are likely to leave the network with a very high accuracy rate compared to the actual data.



Võ Đức Vinh sinh năm 1977, là Kỹ sư Công nghệ Thông tin, Chuyên viên xử lý số liệu, hiện đang công tác tại VNPT Đồng Nai.

Trần Văn Lăng sinh năm 1959, hiện là giảng viên cao cấp của Trường Đại học HUFLIT. Ông tốt nghiệp Khoa Toán Trường Đại học Tổng hợp TP.HCM năm 1982, và nhận học vị tiến sĩ Toán – Lý vào năm 1995. Từ năm 2006 là Phó giáo sư Tin học. Trước khi trở thành giảng viên của HUFLIT, ông là nghiên cứu viên cao cấp của Viện Hàn lâm Khoa học và Công nghệ Việt Nam, thành viên ban lãnh đạo của Viện Cơ học và Tin học ứng dụng, Phó tổng biên tập Tạp chí Tin học và Điều khiển học, Tổng biên tập chuyên mục Điện tử và Viễn thông của Tạp chí Khoa học và Công nghệ Việt Nam. Các lĩnh vực nghiên cứu quan tâm của ông là Tính toán song song và phân tán, Sinh tin học (Bioinformatics), Trí tuệ tính toán (Computational Intelligence), Khoa học và công nghệ tính toán.