HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Đàm Thanh Giang

HỆ THỐNG DỰ ĐOÁN XU HƯỚNG KINH DOANH DỊCH VỤ INTERNET VNPT

Chuyên ngành: Hệ thống thông tin.
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

TP.HCM - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS Tân Hạnh

(Ghi rõ học hàm, học vị)

Phản biện 1:

Phản biện 2:

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: … giờ … ngày … tháng … năm 2022.

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

Đặt vấn đề

Thị trường băng rộng cố định đang ở mức bão hòa, doanh thu tăng trưởng chững lại và việc phát triển thuê bao mới hết sức khó khăn thì chăm sóc và giữ chân khách hàng hiện hữu là hết sức quan trọng, nó không chỉ giúp doanh nghiệp cung cấp dịch vụ phát triển bền vững mà còn ngăn chặn đối thủ phát triển thuê bao mới.

Sự hài lòng của khách hàng khi sử dụng dịch vụ là một trong những nhân tố quan trọng trong việc giữ chân khách hàng. Trong đó việc dự đoán được tập khách hàng có nguy cơ cao rời mạng sẽ giúp cho doanh nghiệp có thể nhanh chóng tiếp cận tư vấn, chăm sóc và đề xuất các gói cước phù hợp là vô cùng quan trọng. Do đó cần có thuật toán dự đoán được tập khách hàng có nguy cơ rời mạng cao nhằm giúp doanh nghiệp kịp thời phản ứng trước các nguy cơ và định hướng phát triển dịch vụ.

Đó là lý do luận văn chọn đề tài: **"Hệ thống dự đoán xu hướng kinh doanh dịch vụ Internet VNPT"**.

Mục đích nghiên cứu

Mục đích nghiên cứu phân tích dữ liệu khách hàng thu thập tại VNPT Tây Ninh:

– Xác định những yếu tố ảnh hưởng đến trải nghiệm sử dụng của khách hàng sử dụng dịch vụ.

– Phân tích và dự đoán để phân tập các nhóm khách hàng có nguy cơ cao, đề xuất các hướng tiếp cận tư vấn và chăm sóc khách hàng.

Đối tượng và phạm vi nghiên cứu

Đối tượng, phạm vi nghiên cứu trên cơ sở dữ liệu thực tế thu thập từ tập khách hàng hiện hữu đang sử dụng dịch vụ Internet của VNPT Tây Ninh.

Nghiên cứu phương pháp xử lý, phân tích dữ liệu, các phương pháp học máy phù hợp với bộ dữ liệu của đề tài, trên nền tảng Python.

Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết:

– Tổng hợp, nghiên cứu các tài liệu về xử lý, mã hóa, phân tích dữ liệu, học máy, kỹ thuật lập trình.

– Sử dụng phương pháp nghiên cứu phân tích dữ liệu, phương pháp dự đoán và phương pháp thực nghiệm để so sánh, đánh giá và phân tích các kết quả đạt được.

Phương pháp nghiên cứu thực nghiệm: sau khi nghiên cứu lý thuyết, tiến hành thực nghiệm kết quả với các phương pháp học máy. Đánh giá các kết quả đạt được; công bố kết quả nghiên cứu.

Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học của luận văn: tập trung phân tích các số liệu thu thập được tại VNPT Tây Ninh, để xác định mức độ tương quan của các yếu tố ảnh hưởng đến trải nghiệm sử dụng dịch vụ của khách hàng.Phân tích các yếu tố ảnh hưởng nhờ áp dụng các phương pháp học máy như LR, SVM, rừng ngẫu nhiên để đưa ra các dự đoán về các tập khách hàng có nguy cơ cao.

Ý nghĩa thực tiễn: xây dựng mô hình dự đoán tập khách hàng có nguy cơ cao để triển khai cho đơn vị tiếp cận

tư vấn chăm sóc, cũng như định hướng được những chính sách ứng phó và phát triển dịch vụ.

Bố cục của báo cáo: báo cáo bao gồm 3 chương cùng với phần mở đầu, phần mục lục, phần kết luận và hướng phát triển, phần tài liệu tham khảo.

Chương 1 – Mô hình hồi quy, các kỹ thuật học máy áp dụng cho bài toán dự đoán.

Chương 2 – Phân tích và đánh giá dữ liệu khách hàng sử dụng dịch vụ FiberVNN của VNPT Tây Ninh.

Chương 3 – Xây dựng mô hình dự đoán tập khách hàng có nguy cơ cao, hỗ trợ đơn vị tiếp cận chăm sóc, cũng như định hướng được những chính sách ứng phó và phát triển dịch vụ. Phân tích và đánh giá kết quả đạt được.

CHƯƠNG 1: MÔ HÌNH HỒI QUY, CÁC KỸ THUẬT HỌC MÁY ÁP DỤNG CHO BÀI TOÁN DỰ ĐOÁN

1.1 Mô hình Logistic Regression

Logistic regression là thuật toán đơn giản nhưng lại rất hiệu quả trong bài toán phân loại (Classification). Logistic regression được áp dụng trong bài toán phân loại nhị phân (Binary classification) tức ta sẽ có hai output, hoặc có thể gọi là hai nhãn (ví dụ như 0 và 1).

1.1.1 Giới thiệu

Logistic Regression (LR) trong phân tích thống kê (hay còn được gọi là mô hình logic) là phân tích hồi quy thích hợp để tiến hành khi biến phụ thuộc là nhị phân (lưỡng phân), nói cách khác là hồi quy với biến phụ thuộc bị giới hạn (Limited Dependent Variable Models).

LR là một mô hình thống kê ở dạng cơ bản của nó sử dụng một hàm logistic để mô hình hóa một biến phụ thuộc nhị phân, mặc dù tồn tại nhiều phần mở rộng phức tạp hơn. Trong phân tích hồi quy, hồi quy logistic (hay hồi quy logic) là ước lượng các tham số của mô hình logistic (một dạng của hồi quy nhị phân). Về mặt toán học, mô hình logistic nhị phân có một biến phụ thuộc với hai giá trị có thể có, chẳng hạn như đạt hoặc không đạt được đại diện bởi một biến chỉ báo, trong đó hai giá trị được gắn nhãn “0” và “1”.

1.1.2 Mô hình Logistic

Xét một mô hình logistic với các tham số cho trước, sau đó xem cách các hệ số có thể được ước tính từ dữ liệu. Hãy xem xét một mô hình có hai yếu tố dự đoán: x_1 và x_2 và một biến nhị phân Bernoulli Y với tham số p = P(Y = 1). Ta giả định mối quan hệ tuyến tính giữa các biến dự đoán và tỷ lệ logic là Y = 1.

Mối quan hệ tuyến tính này có thể được viết ở dạng toán học như sau. Trong đó ℓ là tỷ lệ logic, b là cơ số logarit và β_i là các tham số của mô hình. Ta có:

$$\ell = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Ta có thể khôi phục tỷ lệ logic bằng cách lũy thừa cả hai vế trên:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

Chuyển vế p để ta có xác suất Y = 1:

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}} = S_b(\beta_0+\beta_1 x_1+\beta_2 x_2)$$

Trong đó đẳng thức thứ hai theo sau bằng cách chia tử số và mẫu số của phân số cho $b^{\beta_0+\beta_1 x_1+\beta_2 x_2}$ và trong đó S_b là hàm Sigmoid với cơ số b.

1.1.3 Hàm Sigmoid

Hàm sigmoid là một hàm toán học có đường cong hình chữ "S" hoặc đường cong sigmoid đặc trưng.

1.1.4 Hàm mất mát và phương pháp tối ưu

Hàm logistic là một hàm sigmoid, nhận bất kỳ đầu vào thực tế nào và xuất ra giá trị từ 0 đến 1. [2] Đối với logic, điều này được hiểu là lấy tỷ lệ logic đầu vào và có xác suất đầu ra. Hàm logic tiêu chuẩn: $\sigma: \mathbb{R} \rightarrow (0,1)$ được định nghĩa như sau:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

1.2 Support Vector Machine

SVM (Support Vector Machine) là một thuật toán học máy có giám sát được sử dụng rất phổ biến ngày nay trong các bài toán phân lớp (classification) hay hồi qui (Regression).

Ý tưởng của SVM là tìm một siêu phẳng (hyper lane) để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu.

1.2.1 Giới thiệu

Trong không gian 2 chiều, ta biết rằng khoảng cách từ một điểm có toạ độ (x_0, y_0) tới *đường thẳng* có phương trình $w_1 x + w_2 y + b = 0$ được xác định bởi:

$$\frac{|w_1 x_0 + w_2 y_0 + b|}{\sqrt{w_1^2 + w_2^2}}$$

Trong không gian 3 chiều, khoảng cách từ một điểm có toạ độ (x_0, y_0, z_0) tới một *mặt phẳng* có phương trình $w_1 x + w_2 y + w_3 z + b = 0$ được xác định bởi:

$$\frac{|w_1 x_0 + w_2 y_0 + w_3 z_0 + b|}{\sqrt{w_1^2 + w_2^2 + w_3^2}}$$

Hơn nữa, nếu bỏ trị tuyệt đối ở tử số, có thể xác định được điểm đó nằm về phía nào của *đường thẳng* đang xét. Những điểm làm cho biểu thức trong trị tuyệt đối mang dấu dương nằm về cùng 1, những điểm làm cho biểu thức trong dấu giá trị tuyệt đối mang dấu âm nằm về phía còn lại. Những điểm nằm trên *đường thẳng* sẽ làm cho tử số có giá trị bằng 0, tức khoảng cách bằng 0.

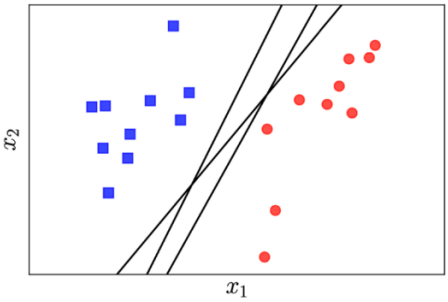
Việc này có thể được tổng quát lên không gian nhiều chiều: Khoảng cách từ một điểm (vector) có toạ độ x_0 tới siêu mặt phẳng (hyperplane) có phương trình $w^T x + b = 0$ được xác định bởi:

$$\frac{w^T x_0 + b}{||w||_2}$$

Với $||w||_2 = \sqrt{\sum_{i=1}^{d} w_i^2}$ với d là số chiều của không gian.

Giả sử rằng có hai lớp khác nhau được mô tả bởi các điểm trong không gian nhiều chiều, hai lớp này *phân tách tuyến tính*, tức tồn tại một siêu phẳng phân chia chính xác hai lớp đó. Hãy tìm một siêu mặt phẳng phân chia hai lớp đó, tức tất cả các điểm thuộc một lớp nằm về cùng một phía của siêu mặt phẳng đó và ngược phía với toàn bộ các điểm thuộc lớp còn lại. Thuật toán Perceptron Learning Algorithm (PLA) [15] có thể làm được việc này nhưng nó có thể cho chúng ta vô số nghiệm như Hình 1.2.

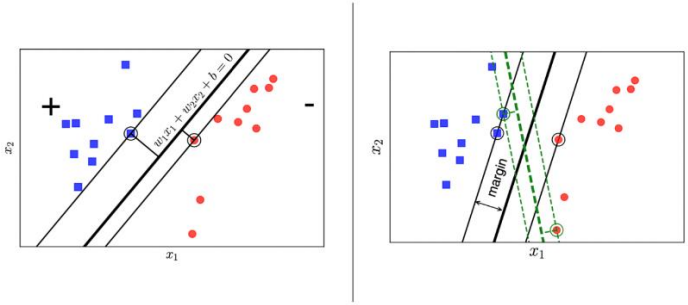
Vấn đề đặt ra là: trong vô số các mặt phân chia, đâu là mặt phân chia tốt nhất *theo một tiêu chuẩn nào đó*? Trong 3 đường thẳng minh họa trong Hình 1.8 phía trên, có hai đường thẳng khá *lệch* về phía lớp hình tròn đỏ. Điều này có thể khiến cho lớp màu đỏ *không thõa mãn bị lấn nhiều quá*. Liệu có cách nào để tìm được đường phân chia mà cả hai lớp đều cảm thõa mãn nhất hay không?



Hình 1.2: Các mặt phân cách hai lớp[1]

1.2.2 Độ rộng của margin

Nếu ta định nghĩa độ thõa mãn của một lớp tỉ lệ thuận với khoảng cách gần nhất từ một điểm của lớp đó tới đường/mặt phân chia, thì ở Hình 1.2 trái, lớp tròn đỏ sẽ không thõa mãn vì đường phân chia gần nó hơn lớp vuông xanh rất nhiều. Chúng ta cần một đường phân chia sao cho khoảng cách từ điểm gần nhất của mỗi lớp (các điểm được khoanh tròn) tới đường phân chia là như nhau. Khoảng cách như nhau này được gọi là *margin*.



Hình 1.3: Margin của hai lớp [1]

Xét tiếp Hình 1.2 bên phải khi khoảng cách từ đường phân chia tới các điểm gần nhất của mỗi lớp là như nhau. Xét hai cách phân chia bởi đường nét liền màu đen và đường nét đứt màu lục, đường nào sẽ làm cho cả hai lớp thõa mãn. Rõ ràng đó phải là đường nét liền màu đen vì nó tạo ra một *margin* rộng hơn.

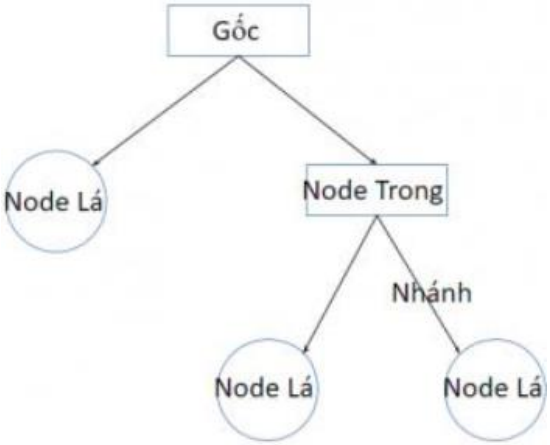
Việc *margin* rộng hơn sẽ mang lại hiệu quả phân lớp tốt hơn vì *sự phân chia giữa hai lớp là rạch ròi hơn*. Bài toán tối ưu trong SVM chính là bài toán đi tìm đường phân chia sao cho *margin* là lớn nhất.

1.3 Thuật toán Cây quyết định

Cây quyết định là một trong những thuật toán máy học phổ biến nhất hiện nay. Nó được dùng trong cả bài toán phân lớp và hồi quy.

1.3.1 Giới thiệu

Cây quyết định là cây mà mỗi nút biểu diễn một đặc trưng (tính chất), mỗi nhánh (branch) biểu diễn một quy luật (rule) và mỗi lá biểu biễn một kết quả (giá trị cụ thể hay một nhánh tiếp tục). [5]



Hình 1.4: Mô hình cây quyết định

Trong cây mô hình quyết định, mỗi nút trung gian [5], tức là nút khác với nút lá và nút gốc, sẽ tương ứng với một phép kiểm tra một thuộc tính. Mỗi nhánh phía dưới của nút đó sẽ tương ứng cho một giá trị của thuộc tính hay còn gọi là kết quả của phép thử. Khác với các nút trung gian, nút lá [5] không chứa thuộc tính cụ thể mà sẽ chứa các nhãn phân lớp. Để xác định nhãn phân lớp cho một dữ liệu mẫu bất kỳ, ta cho dữ liệu mẫu di chuyển từ gốc cây về phía nút lá. Tại mỗi nút trung gian, thuộc tính tương ứng với nút đó được kiểm tra, tùy vào giá trị của thuộc tính đó mà dữ liệu mẫu sẽ được chuyển xuống nhánh bên dưới tương ứng. Quá trình di chuyển này lặp lại cho đến khi dữ liệu mẫu đó tới được nút lá và được gán nhãn phân lớp là nhãn của nút lá tương ứng.

1.3.2 Thuật toán ID3

Thuật toán ID3 được đề ra bởi J. R. Quinlan vào năm 1993 và được sử dụng rộng rãi trong thuật toán cây quyết định. Đây cũng được gọi là thuật toán tham lam (greedy

algorithm) vì thuật toán ID3 tìm kiếm những mô hình mà trong đó các thuộc tính đạt được tối đa lượng thông tin cho việc xác định nhãn lớp của các mẫu trong tập huấn luyện. [11]

Thuật toán ID3 sử dụng Entropy làm cơ sở đo nồng độ đồng nhất của tập dữ liệu.

1.3.3. Thuật toán C4.5

C4.5 là thuật toán dùng để xây dựng cây quyết định được phát triển từ ID3 bởi J. R. Quinlan vào năm 1993. [11] Đặc điểm của C4.5:

– Sử dụng Gain Ratio (thay vì Information Gain) để chọn thuộc tính phân chia trong quá trình dựng cây.

– Xử lý tốt cả hai dạng thuộc tính: rời rạc, liên tục

– Xử lý dữ liệu không đầy đủ (thiếu một số giá trị tại một số thuộc tính).

– C4.5 cho phép các thuộc tính - giá trị bị thiếu có thể thay bằng dấu hỏi (?)

– Những giá trị bị thiếu không được xem xét khi tính toán Information Gain và Gain Radio

– Cắt tỉa cây sau khi xây dựng: Loại bỏ những nhánh cây không thực sự ý nghĩa (thay bằng nút lá).

1.4 Các công trình nghiên cứu trong nước

1.4.1. Áp dụng kỹ thuật khai phá dữ liệu dự báo thuê bao rời mạng trong mạng di động

Luận văn thạc sĩ Công nghệ thông tin "Áp dụng kỹ thuật khai phá dữ liệu dự báo thuê bao rời mạng trong mạng di động" của Nguyễn Ngọc Tuân, Trường Đại học Công nghệ

Hà Nội vào năm 2016. Luận văn đề xuất giải pháp áp dụng khai phá dữ liệu vào bài toán dự báo thuê bao di động rời mạng của Mobifone. Luận văn sử dụng phần mềm mã nguồn mở WEKA để thực nghiệm. [17]

1.4.2. Xây dựng mô hình dự đoán khách hàng tiềm năng cho các gói cước trong mạng di động

Luận văn thạc sĩ Hệ thống thông tin "Xây dựng mô hình dự đoán khách hàng tiềm năng cho các gói cước trong mạng di động" của Đoàn Văn Tâm, Trường Đại học Công nghệ Hà Nội vào năm 2019. Luận văn đề xuất giải pháp sử dụng các kỹ thuật khai phá dữ liệu để dự đoán khách hàng tiềm năng cho các gói cước của tập dữ liệu di động Viettel. Luận văn sử dụng công cụ khai phá dữ liệu Knime để thực nghiệm. [16]

1.5 Các công trình nghiên cứu ngoài nước

1.5.1. Churn Prediction in the Telecommunications Sector Using Support Vector Machines

Ngày nay, với những thách thức do cạnh tranh toàn cầu gây ra, tình trạng mất khách hàng thể hiện là một trong những mối quan tâm đáng kể đối với các công ty trong các ngành công nghiệp khác nhau. Với tỷ lệ tăng trưởng 30%, lĩnh vực viễn thông chiếm vị trí đầu tiên trong danh sách. Để giải quyết vấn đề này, các mô hình dự báo cần được thực hiện để xác định những khách hàng có nguy cơ rời mạng. Trong bài báo này trình bày một phương pháp tiên tiến để dự đoán khách hàng rời mạng trong ngành viễn thông di động. Tập dữ liệu được sử dụng, chứa các bản ghi chi tiết cuộc gọi và có 21 thuộc tính cho mỗi bản ghi trong số 3333 bản ghi của nó. Bài báo sử dụng thuật toán SVM với bốn hàm nhân để triển khai các mô hình dự đoán. Hiệu suất của các mô hình

được đánh giá và so sánh bằng cách sử dụng thước đo độ lợi (gain measure). [3]

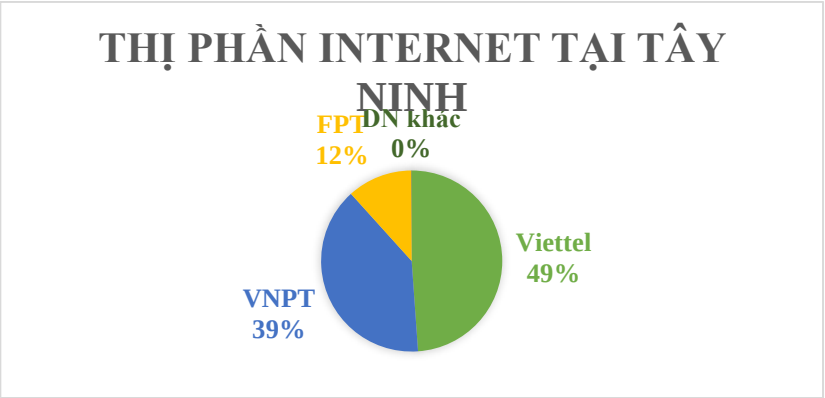
1.5.2. A comparison of machine learning techniques for customer churn prediction

Nghiên cứu so sánh về các phương pháp học máy phổ biến nhất được áp dụng cho vấn đề đầy thách thức về dự đoán chu kỳ khách hàng trong ngành viễn thông. Trong giai đoạn thử nghiệm đầu tiên của nghiên cứu, tất cả các mô hình đã được áp dụng và đánh giá bằng cách sử dụng xác thực chéo trên tập dữ liệu miền công khai, phổ biến. Trong giai đoạn thứ hai, tăng cường và cải thiện hiệu suất. Để xác định các kết hợp tham số hiệu quả nhất, nghiên cứu đã thực hiện một loạt các mô phỏng Monte Carlo cho từng phương pháp và cho một loạt các tham số. Kết quả của nghiên cứu cho thấy sự vượt trội rõ ràng của các phiên bản được tăng cường của các mô hình so với các phiên bản đơn giản (không được tăng cường). Bộ phân loại tổng thể tốt nhất là SVM-POLY sử dụng AdaBoost với độ chính xác gần 97% và độ đo F (F-measure) trên 84%. [4]

CHƯƠNG 2 – PHÂN TÍCH VÀ ĐÁNH GIÁ DỮ LIỆU KHÁCH HÀNG SỬ DỤNG DỊCH VỤ FIBERVNN CỦA VNPT TÂY NINH

1.1. Đánh giá thị trường Internet tại Tây Ninh

Theo dữ liệu thống kê đến cuối năm 2021 trên toàn địa bàn tỉnh Tây Ninh hiện có 3 nhà mạng lớn kinh doanh trong lĩnh vực Internet cáp quang là Viettel, VNPT và FPT. Trong đó, VNPT hiện đang xếp thứ 2 với 39.43% thị phần trên toàn tỉnh.



Hình 2.1: Thị phần Internet tại địa bàn Tây Ninh năm 2021

Với tỷ lệ khách hàng rời mạng so với khách hàng phát triển mới là 34.6%, đây thật sự là gánh nặng cho việc phát triển doanh thu hàng năm của VNPT Tây Ninh. Lý do dẫn đến việc khách hàng rời mạng phụ thuộc vào nhiều yếu tố, trong phần này luận văn sẽ đi sâu phân tích các yếu tố ảnh hưởng trực tiếp đến trải nghiệm sử dụng dịch vụ của khách hàng dẫn đến nguy cơ khách hàng rời mạng.

2.1.1. Các yếu tố về khách hàng

Các yếu tố thuộc về đặc tính của khách hàng gồm:

– Yếu tố vùng miền: Như chúng ta đã biết, mỗi vùng miền sẽ có nhưng đặc trưng riêng, điều kiện kinh tế khác nhau, do đó nhu cầu sử dụng dịch vụ cũng khác nhau, hành vi tiêu dùng cũng khác nhau.

– Loại khách hàng: Những nhóm đối tượng khách hàng khác nhau cũng có nhưng đặc trưng khác nhau, yêu cầu về dịch vụ khác nhau, do đó chắc chắn ảnh hưởng đến nhu cầu sử dụng dịch vụ của khách hàng.

– Thông tin thanh toán của khách hàng: Các hình thức thanh toán khác nhau như: khách hàng đăng ký gói chu kỳ dài hay trả hàng tháng cũng ảnh hưởng đến trải nghiệm của khách hàng. Khách hàng đăng ký gói chu kỳ dài sẽ ít vướng mắc vào vấn đề cước và nợ cước nên sẽ có trải nghiệm dịch vụ tốt hơn.

2.1.2. Các yếu tố về chất lượng dịch vụ

Các yếu tố chất lượng dịch vụ là chất lượng của từng dịch vụ cung cấp bao gồm:

– Băng thông: là băng thông tối đa của một gói cước khi cung cấp cho khách hàng.

– Tích hợp gói cước: tùy vào nhu cầu sử dụng của khách hàng, khách hàng có thể hưởng được những ưu đãi nhất định khi đăng ký tích hợp nhiều dịch vụ như: di động, băng rộng cố định, truyền hình MyTV…

– Tình trạng suy hao: Do chất lượng thiết bị không tốt, các mối nối không được thực hiện đúng kỹ thuật… gây nên tình trạng suy hao tín hiệu, dẫn đến chất lượng dịch vụ bị suy giảm.

– Thời gian ngắt quãng dịch vụ: do các vấn đề về cước và nợ cước hoặc do các yếu tố khách quan khác dẫn đến dịch vụ của khách hàng bị ngắt quãng.

1.2. Bài toán chăm sóc và dự đoán khách hàng rời mạng của VNPT Tây Ninh

Dựa theo dữ liệu trên hệ thống quản trị của Tây Ninh, lý do thuê bao Internet cáp quang rời mạng như sau:

– 1.26% trường hợp do sự lôi kéo của đối thủ cạnh tranh (Đối thủ kéo cáp vào nhà cho khách hàng dùng thử miễn phí, chính sách hấp dẫn hơn…);

– 3.21% trường hợp do chất lượng phục vụ và dịch vụ kém hoặc thiết bị đầu cuối kém, sửa chữa nhiều lần chưa khắc phục được;

– 16.22% do yếu tố khách quan khác như: khu vực bị giải tỏa, chuyển nhà, khách hàng chỉ sử dụng dịch vụ trong thời gian ngắn (do thuê nhà, hợp đồng thời vụ tại các khu công nghiệp), thi công, sửa nhà …

– 5.56% do khách hàng không có nhu cầu nữa (Thừa đường truyền Internet, không quản lý được con cái, chuyển sang sử dụng 3G,4G…)

– Còn lại 73.74% do khóa nợ cước. Tuy nhiên, đây không phải là nguyên nhân thật sự, mà chỉ là kết quả. Bị khóa do nợ cước, có thể do khách hàng đã bị đối thủ lôi kéo, chất lượng dịch vụ, thiết bị đầu cuối kém, sửa chữa nhiều lần, không còn nhu cầu, thái độ phục vụ…từ đó khách hàng không thanh toán cước.

Hiện nay, đối với vấn đề giám sát và theo dõi chăm sóc khách hàng tại VNPT Tây Ninh được thực hiện dựa hoàn toàn vào yếu tố con người, tại tất cả các điểm chạm như: nhân viên thu cước, nhân viên kỹ thuật, nhân viên

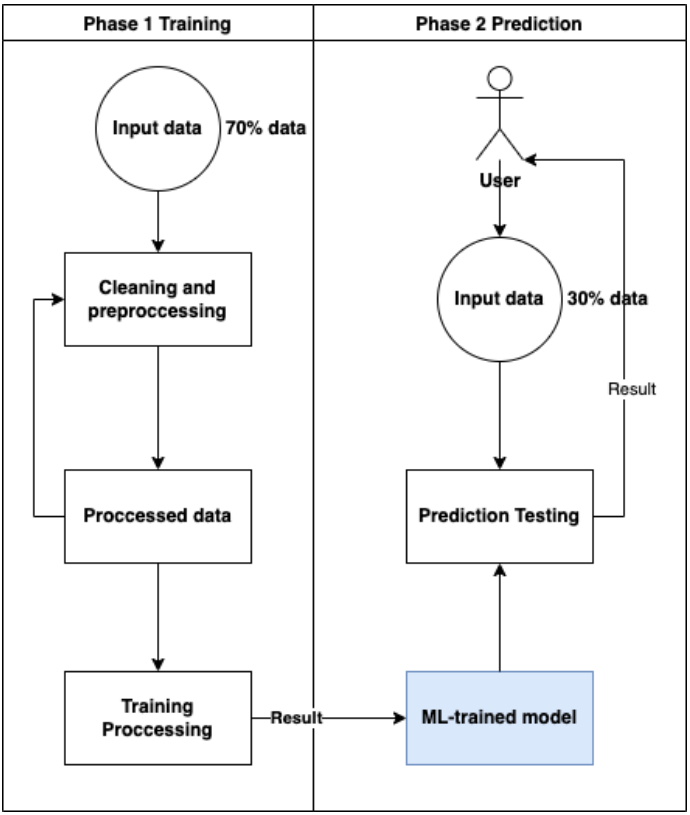
quản lý địa bàn... và được điều hành bởi cấp lãnh đạo Phòng Bán hàng, Trung tâm Kinh doanh và Viễn thông tỉnh. Điều này đòi hỏi rất nhiều vào các yếu tố con người, từ kỹ năng của nhân viên cho đến năng lực điều phối, đôn đốc và giám sát của các cấp Lãnh đạo.

Đối với dữ liệu phân tích, hiện nay VNPT Tây Ninh chú trọng vào việc chăm sóc khách hàng có thực hiện các cuộc gọi báo hỏng, các khách hàng không phát sinh lưu lượng 5 ngày, cũng như dựa vào tình trạng khóa và nợ cước của từng thuê bao. Việc giám đánh giá các yếu tố rời mạng chỉ được thực hiện sau khi khách hàng rời mạng và do nhân viên nhập các lý do của từng khách hàng lên hệ thống điều hành kinh doanh tại đơn vị. Điều này dẫn đến việc phân tích chưa thật sự chính xác và khách quan để phản ánh tình hình thực tế từ phía khách hàng.

Từ đó, đề tài nghiên cứu áp dụng các kỹ thuật học máy vào việc dự đoán nguy cơ khách hàng sử dụng Internet cáp quang rời mạng và tiến hành đánh giá kết quả thực nghiệm tại VNPT Tây Ninh, đưa ra hướng phát triển mở rộng của để tài để đáp ứng những nhu cầu triển khai thực tế tại đơn vị.

CHƯƠNG 3 - XÂY DỰNG MÔ HÌNH

Quá trình để xây dựng mô hình dự đoán dữ liệu khách hàng có nguy cơ cao được mô phỏng theo hình 3.1.



Hình 3.1: Mô tả quy trình dự đoán

3.1. Chuẩn bị và tiền xử lý dữ liệu

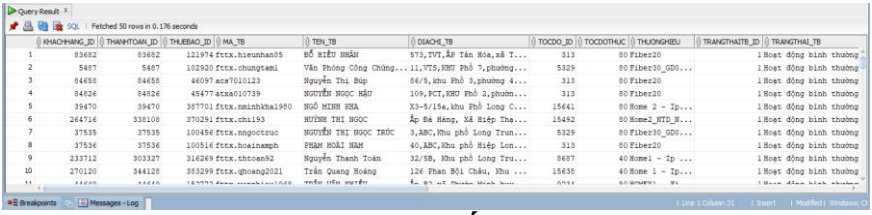
Giai đoạn chuẩn bị và xử lý dữ liệu ban đầu luôn là một giai đoạn quan trọng trong quy trình khai phá dữ liệu. Dữ liệu là một trong hai thành phần của phân lớp dữ liệu. Thông tin khách hàng cần thiết để dự đoán thuê bao rời mạng bao gồm: thông tin về khách hàng, thông tin về thanh

toán, thông tin quá trình sử dụng…. Từ các dữ liệu khác nhau, một cơ sở dữ liệu đưa ra dự đoán về việc rời khỏi mạng được xây dựng với dữ liệu thu thập mục tiêu.

Dữ liệu thu thập được sau khi lọc và xử lý có 102,061 dòng gồm các thông tin:

Dữ liệu khách hàng: ngày bắt đầu sử dụng dịch vụ, doanh thu, số tháng nợ cước, số tiền nợ cước, hình thức thanh toán dịch vụ, mức độ hài lòng khi sử dụng dịch vụ, số lần báo hỏng trong khoảng thời gian sử dụng dịch vụ.

Dữ liệu về chất lượng dịch vụ: băng thông, gói tích hợp, số ngày cắt dịch vụ, số ngày sử dụng dịch vụ.



Hình 3.2: Dữ liệu thực tế Oracle tại Tây Ninh

Dữ liệu thu thập là dữ liệu Internet cáp quang tại đơn vị tổng hợp từ nhiều nguồn trong 03 tháng tháng từ tháng 06/2021 đến 12/2021. Chi tiết các trường dữ liệu được mô tả trong Bảng 3.1.

Bảng 3.1: Mô tả dữ liệu Internet cáp quang của VNPT Tây Ninh

STT	Trường dữ liệu	Mô tả	Kiểu dữ liệu
1	THUEBAO_ID	Thuê bao ID là duy nhất cho mỗi thuê bao	Số nguyên

2	MA_TB	Mã thuê bao là duy nhất cho mỗi thuê bao	Chuỗi
3	TEN_TB	Tên thuê bao	Chuỗi
4	DIACHI_TB	Địa chỉ thuê bao	Chuỗi
5	TOCDOTHUC	Tốc độ Internet	Số nguyên
6	NGAY_SD	Thời gian bắt đầu sử dụng dịch vụ	Ngày giờ
7	NGAY_TD	Ngày khóa/tạm ngưng dịch vụ	Ngày giờ
8	NGAY_CAT	Ngày hủy dịch vụ	Ngày giờ
9	SOTHANG_SD	Số tháng sử dụng dịch vụ tính đến 31/12/2021 hoặc tính đến ngày hủy dịch vụ.	Số nguyên
10	SONGAY_SD	Số ngày sử dụng dịch vụ trong 06 tháng gần nhất	Số nguyên
11	SONGAY_KHOA	Số ngày dịch vụ bị ngắt quãng trong 06 tháng gần nhất	Số nguyên
12	NGAYTHANG	Tổng số ngày 06 tháng gần nhất	Số nguyên
13	TRATRUOC	Hình thức thanh toán của khách hàng.	1: Khách hàng đăng ký

			gói dài ngày 0: Khách hàng trả hàng tháng
14	DOANHTHU	Doanh thu phát sinh	Số nguyên
15	TIENNO	Tiền nợ của khách hàng tính đến 30/11/2021	Số nguyên
16	SOTHANG_NO	Tổng số tháng nợ của khách hàng	Số nguyên
17	GOI_DADV	Khách hàng sử dụng gói tích hợp nhiều dịch vụ hoặc riêng lẻ	1: Tích hợp 0: Riêng lẻ
18	SOLAN_BAOHONG	Số lần báo hỏng của khách hàng trong 06 tháng gần nhất	Số nguyên
19	TONG_KHAOSAT	Số lần khảo sát mức độ hài lòng của khách hàng trong 06 tháng gần nhất	Số nguyên
20	SOLAN_HAILONG	Tổng số lần hài lòng của khách hàng	Số nguyên

21	SOLAN_KO_HAI LONG	Tổng số lần không hài lòng của khách hàng	Số nguyên
22	LOAI_KH	Loại khách hàng là Khách hàng Cá nhân hoặc Khách hàng doanh nghiệp	1: KHDN 0: KHCN
23	TEN_QUAN	Huyện/Thành phố	Chuỗi
24	TEN_PHUONG	Phường/xã	Chuỗi
25	LOAI_KV	Xếp loại khu vực	1: Khu vực loại 1 2: Khu vực loại 2 3: Khu vực loại 3
26	KHONG_PSLL	Thuê bao 5 ngày không phát sinh lưu lượng	Có: 1 Không: 0
27	TRANGTHAITB_ ID	Trạng thái của thuê bao sử dụng dịch vụ	1: Hoạt động bình thường 5: Khóa 2 chiều do nợ cước 6: Tạm ngưng

			theo yêu cầu 7: Thanh lý theo yêu cầu 9: Thanh lý cưỡng bức
28	ROIMANG	Khách hàng có rời mạng hay không	1: Rời mạng 0: Không rời mạng

Từ bảng dữ liệu 3.1 tiến hành làm sạch dữ liệu bằng cách loại bỏ các dòng dữ liệu có trường trống hoặc NULL, các trường dữ liệu bất thường như nợ ghi nhận âm… Loại bỏ một số trường mang tính bảo mật người dùng: họ tên, địa chỉ, mã thuê bao... Tiến hành chuyển đổi kiểu dữ liệu từ dạng chữ (chuỗi) sang dạng số bằng cách mã hóa các kí tự bằng số.

Bảng 3.2: Mô tả dữ liệu sau khi thực hiện làm sạch

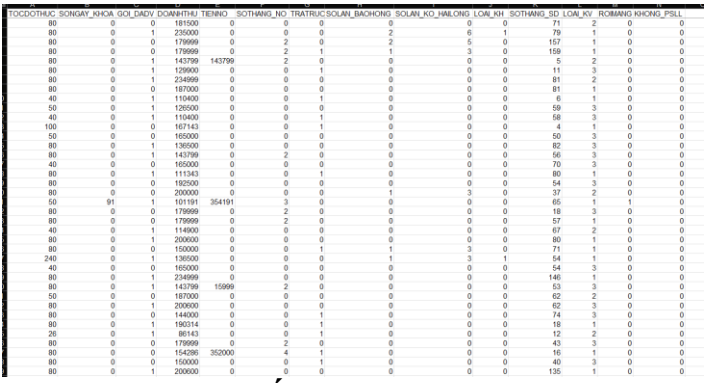
STT	Trường dữ liệu	Mô tả	Kiểu dữ liệu
1	THUEBAO_ID	Thuê bao ID là duy nhất cho	Số nguyên

		mỗi thuê bao	
5	TOCDOTHUC	Tốc độ Internet	Số nguyên
9	SOTHANG_SD	Số tháng sử dụng dịch vụ tính đến 31/12/2021 hoặc tính đến ngày hủy dịch vụ.	Số nguyên
11	SONGAY_KHOA	Số ngày dịch vụ bị ngắt quãng trong 06 tháng gần nhất	Số nguyên
13	TRATRUOC	Hình thức thanh toán của khách hàng.	1: Khách hàng đăng ký gói dài ngày 0: Khách hàng trả hàng tháng

14	DOANHTHU	Doanh thu phát sinh	Số nguyên
15	TIENNO	Tiền nợ của khách hàng tính đến 30/11/2021	Số nguyên
16	SOTHANG_NO	Tổng số tháng nợ của khách hàng	Số nguyên
17	GOI_DADV	Khách hàng sử dụng gói tích hợp nhiều dịch vụ hoặc riêng lẻ	1: Tích hợp 0: Riêng lẻ
18	SOLAN_BAOHONG	Số lần báo hỏng của khách hàng trong 06 tháng gần nhất	Số nguyên
21	SOLAN_KO_HAILONG	Tổng số lần không hài lòng của khách hàng	Số nguyên
22	LOAI_KH	Loại khách hàng là	1: KHDN

		Khách hàng Cá nhân hoặc Khách hàng doanh nghiệp	0: KHCN
25	LOAI_KV	Xếp loại khu vực	1: Khu vực loại 1 2: Khu vực loại 2 3: Khu vực loại 3
26	KHONG_PSLL	Thuê bao 5 ngày không phát sinh lưu lượng	Có: 1 Không: 0
27	ROIMANG	Khách hàng có rời mạng hay không	1: Rời mạng 0: Không rời mạng

Thu được kết quả ở dạng mã hóa như sau:



TOCDOTHUC	SONGAY_KHOA	GOI_DADV	DOANHTHU	TIENNO	SOTHANG_NO	TRATRUC	SOLAN_BAOHONG	SOLAN_KO_HAILONG	LOAI_KH	SOTHANG_SD	LOAI_KV	ROIMANG	KHONG_PSLL
80	0	0	181500	0	0	0	0	0	0	71	2	0	0
80	0	1	235000	0	0	0	2	6	1	79	1	0	0
80	0	0	179999	0	2	0	2	5	0	157	1	0	0
80	0	0	179999	0	2	1	1	3	0	159	1	0	0
80	0	1	143799	143799	2	0	0	0	0	5	2	0	0
80	0	1	129900	0	0	1	0	0	0	11	3	0	0
80	0	1	234999	0	0	0	0	0	0	81	2	0	0
80	0	0	187000	0	0	0	0	0	0	81	1	0	0
40	0	1	110400	0	0	1	0	0	0	6	1	0	0
50	0	1	126500	0	0	0	0	0	0	59	3	0	0
40	0	1	110400	0	0	1	0	0	0	58	3	0	0
100	0	0	167143	0	0	1	0	0	0	4	1	0	0
50	0	0	165000	0	0	0	0	0	0	50	3	0	0
80	0	1	136500	0	0	0	0	0	0	82	3	0	0
80	0	1	143799	0	2	0	0	0	0	56	3	0	0
40	0	0	165000	0	0	0	0	0	0	70	3	0	0
80	0	1	111343	0	0	1	0	0	0	80	1	0	0
80	0	0	192500	0	0	0	0	0	0	54	3	0	0
80	0	0	200000	0	0	0	1	3	0	37	2	0	0
50	91	1	101191	354191	3	0	0	0	0	65	1	1	0
80	0	0	179999	0	2	0	0	0	0	18	3	0	0
80	0	0	179999	0	2	0	0	0	0	57	1	0	0
40	0	1	114900	0	0	0	0	0	0	67	2	0	0
80	0	1	200600	0	0	0	0	0	0	80	1	0	0
80	0	0	150000	0	0	1	1	3	0	71	1	0	0
240	0	1	136500	0	0	0	1	3	1	54	1	0	0
40	0	0	165000	0	0	0	0	0	0	54	3	0	0
80	0	1	234999	0	0	0	0	0	0	146	1	0	0
80	0	1	143799	15999	2	0	0	0	0	53	3	0	0
50	0	0	187000	0	0	0	0	0	0	62	2	0	0
80	0	1	200600	0	0	0	0	0	0	62	3	0	0
80	0	0	144000	0	0	1	0	0	0	74	3	0	0
80	0	1	190314	0	0	1	0	0	0	18	1	0	0
26	0	1	86143	0	0	1	0	0	0	12	2	0	0
80	0	0	179999	0	2	0	0	0	0	43	3	0	0
80	0	0	154286	352000	4	1	0	0	0	16	1	0	0
80	0	0	150000	0	0	1	0	0	0	40	3	0	0
80	0	1	200600	0	0	0	0	0	0	135	1	0	0

Hình 3.3: Kết quả làm sạch dữ liệu

Sử dụng hàm fit_stransform để chuyển đổi dữ liệu dạng số nguyên như: tốc độ, số ngày khóa, doanh thu, tiền nợ, số tháng nợ, số lần báo hỏng, số lần không hài lòng, số tháng sử dụng.

```
[ ]  #feature scaling
     from sklearn.preprocessing import MinMaxScaler
     sc = MinMaxScaler()
     data_df['TOCDOTHUC'] = sc.fit_transform(data_df[['TOCDOTHUC']])
     data_df['SONGAY_KHOA'] = sc.fit_transform(data_df[['SONGAY_KHOA']])
     data_df['DOANHTHU'] = sc.fit_transform(data_df[['DOANHTHU']])
     data_df['TIENNO'] = sc.fit_transform(data_df[['TIENNO']])
     data_df['SOTHANG_NO'] = sc.fit_transform(data_df[['SOTHANG_NO']])
     data_df['SOLAN_BAOHONG'] = sc.fit_transform(data_df[['SOLAN_BAOHONG']])
     data_df['SOLAN_KO_HAILONG'] = sc.fit_transform(data_df[['SOLAN_KO_HAILONG']])
     data_df['SOTHANG_SD'] = sc.fit_transform(data_df[['SOTHANG_SD']])
```

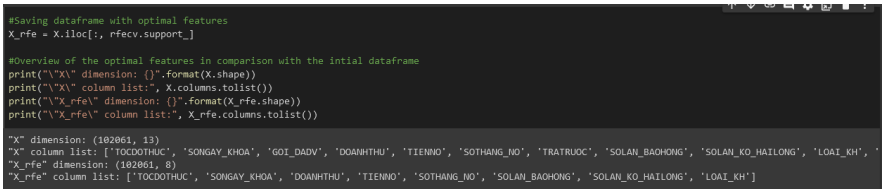
Hình 3.4: Scaling dữ liệu

Sử dụng thư viện RFECV để tính độ tương quan của các trường dữ liệu trong tập dữ liệu hiện tại.

```
# Feature selection to improve model building
from sklearn.feature_selection import RFECV
from sklearn.model_selection import StratifiedKFold
log = LogisticRegression()
rfecv = RFECV(estimator=log, cv=StratifiedKFold(10, random_state=50, shuffle=True), scoring="accuracy")
rfecv.fit(X, y)
```

Hình 3.5: Tính toán mức độ tương quan của các trường dữ liệu

Kết quả ta thu được số lượng trường dữ liệu được lựa chọn là 8 bao gồm: Tốc độ, Số ngày khóa, Doanh thu, Tiền nợ, Số tháng nợ, Số lần báo hỏng, Số lần không hài lòng, Loại khách hàng.



```
#Saving dataframe with optimal features
X_rfe = X.iloc[:, rfecv.support_]

#Overview of the optimal features in comparison with the intial dataframe
print("\"X\" dimension: {}".format(X.shape))
print("\"X\" column list:", X.columns.tolist())
print("\"X_rfe\" dimension: {}".format(X_rfe.shape))
print("\"X_rfe\" column list:", X_rfe.columns.tolist())
```

```
"X" dimension: (102061, 13)
"X" column list: ['TOCDOTHUC', 'SONGAY_KHOA', 'GOI_DADV', 'DOANHTHU', 'TIENNO', 'SOTHANG_NO', 'TRATRUOC', 'SOLAN_BAOHONG', 'SOLAN_KO_HAILONG', 'LOAI_KH', '
"X_rfe" dimension: (102061, 8)
"X_rfe" column list: ['TOCDOTHUC', 'SONGAY_KHOA', 'DOANHTHU', 'TIENNO', 'SOTHANG_NO', 'SOLAN_BAOHONG', 'SOLAN_KO_HAILONG', 'LOAI_KH']
```

Hình 3.6: Các trường dữ liệu được lựa chọn

3.2. Thư viện Scikit-learn

Scikit-learn (Sklearn) là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling gồm: classification, regression, clustering, và dimensionality reduction.

Thư viện được cấp phép bản quyền chuẩn FreeBSD và chạy được trên nhiều nền tảng Linux. Scikit-learn được sử dụng như một tài liệu để học tập.

3.3. Tiến hành thực hiện dự đoán dữ liệu

Thực nghiệm dữ liệu thực tế với các mô hình dự báo như: Logistic Regression Classification, SVM Classification, Random Forest Classification, Decision Tree Classification, Naive Bayes Classification. Thu thập, đánh giá kết quả và lựa chọn mô hình tối ưu.

3.3.1 Dự đoán bằng mô hình LR

Bảng 3.4: Kết quả dự đoán bằng mô hình LR

		Lớp dự đoán (predicted class)	
		Đúng	**Sai**
Lớp thực tế (actual class)	**Đúng**	TP = 30000	FN = 74
	Sai	FP = 231	TN = 596
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.9901$			

3.3.2 Dự đoán bằng SVM

Bảng 3.5: Kết quả dự đoán bằng SVM

		Lớp dự đoán (predicted class)	
		Đúng	**Sai**
Lớp thực tế (actual class)	**Đúng**	TP = 30000	FN = 100
	Sai	FP = 108	TN = 719
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.9933$			

3.3.3 Dự đoán bằng Random Forest

Bảng 3.6: Kết quả dự đoán bằng Random Forest

		Lớp dự đoán (predicted class)	
		Đúng	**Sai**
Lớp thực tế (actual class)	**Đúng**	TP = 30000	FN = 89
	Sai	FP = 97	TN = 730
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.9940$			

3.3.4 Dự đoán bằng Decision Tree

Bảng 3.7: Kết quả dự đoán bằng Decision Tree

		Lớp dự đoán (predicted class)	
		Đúng	**Sai**
Lớp thực tế (actual class)	**Đúng**	TP = 30000	FN = 98
	Sai	FP = 128	TN = 699
$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.9927$			

3.1. Kết quả dự đoán và đánh giá

3.4.1 Độ chính xác của thuật toán

Cách đơn giản và hay được sử dụng nhất là accuracy (độ chính xác). Cách đánh giá này đơn giản tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

Giả sử ta có bài toán phân lớp với đầu ra là 2 lớp Đúng/Sai, kết quả phân lớp trên tập mẫu so với thực tế có 4 khả năng thể hiện… Bảng này được gọi là ma trận sai số (confusion matrix).

Bảng 3.8: Bảng ma trận sai số

		Lớp dự đoán (predicted class)	
		Đúng	**Sai**
Lớp thực tế (actual class)	**Đúng**	True Positive (TP)	False Negative (FN)
	Sai	False Positive (FP)	True Negative (TN)

True Positive thể hiện khả năng dự đoán phân lớp phân lớp đúng của một mẫu dự đoán thuộc phân lớp Đúng, False Positive thể hiện khả năng dự đoán phân lớp sai của một mẫu được dự đoán thuộc phân lớp Đúng.

False Negative thể hiện khả năng dự đoán phân lớp phân lớp đúng của một mẫu dự đoán thuộc phân lớp Sai, True Negative thể hiện khả năng dự đoán phân lớp sai của một mẫu được dự đoán thuộc phân lớp Sai.

Ta có độ đo đánh giá hiệu quả của kết quả phân lớp như sau:

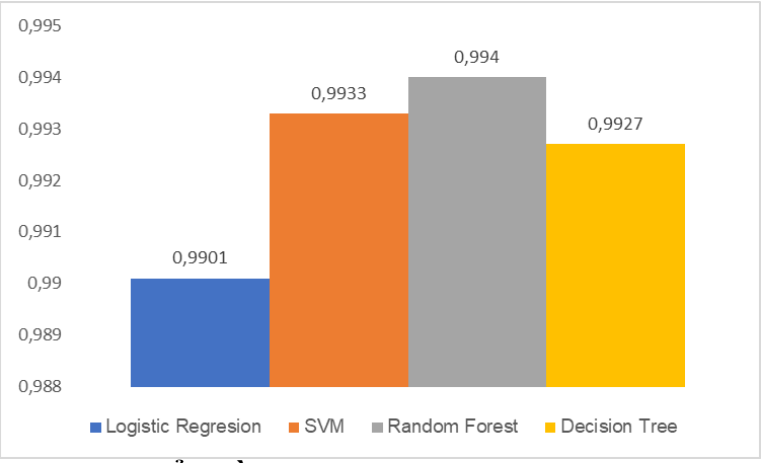
Bảng 3.9: Cách tính độ chính xác

Tên độ đo	Công thức	Diễn giải
Độ chính xác	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	Tỷ lệ các mẫu được phân lớp đúng trên toàn bộ tập mẫu

3.4.2 Kết quả dự đoán và đánh giá

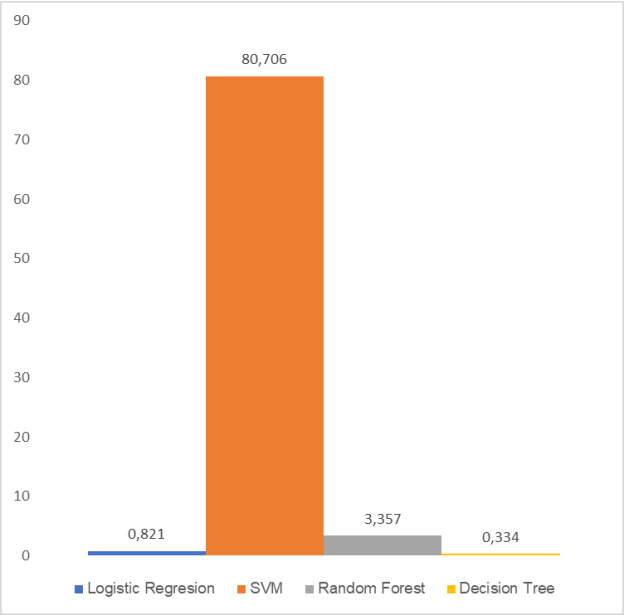
Bảng 3.10: Kết quả dự đoán của các mô hình

Mô hình	Độ chính xác	Thời gian(s)
Logistic Regression	0.9901	0.821
SVM	0.9933	80.706
Random Forest	0.9940	3.357
Decision Tree	0.9927	0.334



Hình 3.7: Biểu đồ so sánh độ chính xác của các thuật toán phân lớp

Bảng 3.7 là kết quả dự đoán của các mô hình dựa trên các độ đo được trình bày trong mục 3.1. Từ kết quả dự đoán có thể thấy các mô hình cho kết quả sấp xỉ không chênh lệch quá nhiều. Trong đó, mô hình sử dụng Random Forest cho kết quả tốt nhất trên cùng một tập dữ liệu so với các mô hình còn lại.



Hình 3.8: Biểu đồ so sánh thời gian huấn luyện của các thuật toán phân lớp (đơn vị giây)

Qua đó ta nhận thấy mô hình sử dụng thuật toán RF cho kết quả tối ưu nhất về độ chính xác, còn thuật toán DT cho kết tối ưu nhất về thời gian thực thi. Do đó giải quyết bài toán dự đoán số khách hàng rời mạng theo từng tháng, quý hoặc năm cho tập dữ liệu Internet cáp quang tại VNPT Tây Ninh, ta có thể tiến hành như sau:

– Đối với tập dữ liệu khách hàng lớn, đòi hỏi phải tối ưu thời gian thực thi, chúng ta sẽ áp dụng thuật toán Decision Tree cho bài toán dự đoán tập khách hàng có nguy cơ cao.

– Đối với tập khách hàng vừa và nhỏ, đòi hỏi phải tối ưu về độ chính xác, chúng ta sẽ áp dụng thuật toán Random Forest cho bài toán dự đoán tập khách hàng có nguy cơ cao.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết quả đạt được

Thông qua đề tài nghiên cứu, luận văn đã đề xuất và thực nghiệm được mô hình dự đoán tập khách hàng có nguy cơ cao dựa trên dữ liệu người dùng thực tế. Từ đó giúp cho đơn vị chủ động được trong công tác chăm sóc khách hàng và định hướng phát triển dịch vụ.

1.1. Về mặt lý thuyết

Khai thác được mô hình dữ liệu khách hàng có nguy cơ cao để xây dựng mô hình phát hiện và cảnh báo nguy cơ rời mạng.

Ứng dụng Trí tuệ nhân tạo (AI), Machine Learning, các thuật toán học máy và phương pháp khai phá dữ liệu vào việc phát hiện khách hàng có nguy cơ cao.

Khai thác được các thuật toán phân lớp dữ liệu, cụ thể là các mô hình LR, SVM, RF, Cây quyết định… Thực nghiệm trên ứng dụng thực tế, thu thập kết quả và đánh giá thuật toán tối ưu cho bài toán.

Ứng dụng thư viện scikit-learn trên nền tảng python vào việc nghiên cứu các vấn đề học máy, sử dụng được các tham số để tối ưu mô hình dự đoán.

1.2. Về mặt thực tiễn

Luận văn đã đưa ra được giải pháp phát hiện khách hàng có nguy cơ cao và cảnh báo sớm cho đơn vị dựa vào các dữ liệu lưu trữ trên hệ thống. Việc này sẽ làm tiền đề để xây dựng một công cụ cảnh báo khách hàng có nguy cơ cao phục vụ cho việc chăm sóc lôi kéo khách hàng trong tương lai, thay thế cho công tác đang vận hành nhân công tại đơn vị.

Xây dựng mô hình dự đoán khách hàng nguy cơ cao, phân tích và đánh giá mô hình xây dựng được để hiểu rõ hơn về cách thức hoạt động của các thuật toán khai phá dữ liệu.

2. Hạn chế

Do dữ liệu thực tế có sự chênh lệch khá lớn giữa số lượng thuê bao thanh lý và số lượng thuê bao hiện hữu, dẫn đến kết quả của mô hình chưa cao cũng như chưa bao quát được hết các trường hợp. Dữ liệu mẫu cần được training và mở rộng môi trường áp dụng.

Các trường hợp phân loại sai vẫn còn nhiều dẫn đến việc nhắm mục tiêu khách hàng có nguy cơ cao chưa thật sự chuẩn xác.

Mô hình dự đoán trong luận văn còn ở mức cơ bản, chưa phân tích sâu vào các tham số để phù hợp với mô hình dữ liệu thực tế.

3. Hướng phát triển

Tập trung nghiên cứu rút trích các đặc trưng thuộc tính phù hợp hơn cho quá trình phân tích, tăng độ chính xác trong việc dự đoán tập khách hàng có nguy cơ cao. Nghiên cứu các mô hình dự đoán để cải thiện mô hình dự đoán được tốt hơn.

Nghiên cứu áp dụng các mô hình phân loại kết hợp để tìm kiếm những mô hình tối ưu phù hợp với dữ liệu thực tế tại đơn vị.

Tiến hành áp dụng tại VNPT Tây Ninh. Cảnh báo sớm các nhóm khách hàng có nguy cơ cao, góp phần hỗ trợ công tác chăm sóc và lôi kéo khách hàng được tiến hành nhanh và hiệu quả hơn. Từ đó, góp phần thúc đẩy hiệu quả kinh doanh tại đơn vị.