HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Lê Đức Hòa Bình

HỆ HỖ TRỢ QUYẾT ĐỊNH KINH DOANH DỊCH VỤ VIỄN THÔNG THEO XU HƯỚNG KHÁCH HÀNG Ở TÂY NINH

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**

Phản biện 1: PGS. TS. Trần Mạnh Hà.

Phản biện 2: PGS. TS. Thoại Nam.

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: 8 giờ 00 ngày 02 tháng 07 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

Việc khách hàng hài lòng sau khi sử dụng dịch vụ phụ thuộc vào rất nhiều yếu tố khách quan và chủ quan. Trong đó tư vấn cho khách hàng một gói cước phù hợp là cực kì quan trọng. Việc này lâu nay vẫn thường xuyên được phân tích, tuy nhiên thực hiện bằng các biện pháp thủ công, thô sơ mất rất nhiều thời gian, và đòi hỏi người phân tích phải có chuyên môn tương đối tốt, nhưng độ chính xác mang lại tương đối không cao.

Do đó để có biện pháp phấn tích khoa học và hiện đại khắc phục các tồn tại như đã mô tả, khi đề tài hoàn thiện nhiều người có thể sử dụng. Trong báo cáo này sử phương pháp học máy để phân tích dự đoán các yếu tố ảnh hưởng đến gói cước sử dụng dịch vụ của khách hàng tại VNPT Tây Ninh. Kết quả tư vấn chính xác, nhanh giúp doanh nghiệp phát triển khách hàng mới, cũng như đảm bảo chất lượng dịch vụ phù hợp với nhu cầu sử dụng của khách hàng.

Mục đích nghiên cứu

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

- Xác định các yếu tố có ảnh hưởng đến gói cước phù hợp nhất với khách hàng.

- Phân tích sự ảnh hưởng của các yếu tố đó như thế nào đến gói cước mà khách hàng cần đăng ký.

- Đề xuất gói cước cho khách hàng bằng học máy.

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 nên 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.

Bố cục của báo cáo: báo cáo bao gồm 5 chương cùng với phần mở đầu, phần mục lục, phần tài liệu tham khảo.

Chương 1- Cơ sở lý thuyết và các công trình nghiên cứu có liên quan: Trình bày một số khải niệm có liên quan đến máy học, thuật toán cây quyết định. Ngoài ra, chương 1 còn đề cập đến một số công trình nghiên cứu có liên quan.

Chương 2 – Cây quyết định, Rừng ngẫu nhiên: Trình bày về bài toán phân lớp, Cây quyết định, Rừng ngẫu nhiên và thư viện Scikit Learn.

Chương 3 - Xây dựng mô hình: Trình bày các bước xây dựng mô hình khuyến nghị gói cước dựa vào thuật toán Rừng ngẫu nhiên.

Chương 4 – Phân tích và đánh giá: Đánh giá kết quả đạt được sau khi xây dựng mô hình Khuyến nghị gói cước dựa vào mức độ chính xác của mô hình.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU CÓ LIÊN QUAN

1.1. Giới thiệu bài toán phân lớp dữ liệu và các vấn đề liên quan

1.1.1. Khái niệm về phân lớp dữ liệu và bài toán phân lớp dữ liệu

Phân lớp (classification) dữ liệu là một tiến trình xử lý nhằm xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước. Các mẫu dữ liệu hay các đối tượng được xếp vào các lớp dựa trên giá trị của các thuộc tính (attributes) của mẫu dữ liệu hay đối tượng. Quá trình phân lớp dữ liệu kết thúc khi tất cả các dữ liệu đã được xếp vào các lớp tương ứng. Khi đó, mỗi lớp dữ liệu được đặc trưng bởi tập các thuộc tính của các đối tượng chứa trong lớp đó.

Quy trình giải quyết bài toán phân lớp dữ liệu

(1) Giai đoạn huấn luyện

(2) Giai đoạn kiểm chứng

1.1.2. Các độ đo đánh giá mô hình phân lớp dữ liệu

(1) Độ đo Precision (Mức chính xác)

- Định nghĩa: Precision = TP / (TP + FP).

- Ý nghĩa: Giá trị Precision càng cao thể hiện khả năng càng cao để một kết quả phân lớp dữ liệu được đưa ra bởi bộ phân lớp là chính xác.

(2) Độ đo Recall (Độ bao phủ, độ nhạy hoặc độ triệu hồi)

- Định nghĩa: Recall = TP / (TP + FN).

- Ý nghĩa: Giá trị Recall càng cao thể hiện khả năng kết quả đúng trong số các kết quả đưa ra của bộ phân lớp càng cao.

(3) Độ đo Accuracy (Độ chính xác)

- Định nghĩa: Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100%.

- Ý nghĩa: Accuracy phản ánh độ chính xác chung của bộ phân lớp dữ liệu..

(4) Độ đo F-Measure

- Định nghĩa: F-Measure = 2.(Precision.Recall) / (Precision + Recall).

- Ý nghĩa: F-Measure là độ đo nhằm đánh giá độ chính xác thông qua quá trình kiểm chứng dựa trên sự xem xét đến hai độ đo là Precision và Recall. Giá trị F-Measure càng cao phản ánh độ chính xác càng cao của bộ phân lớp dữ liệu. Có thể coi độ đo F-Measure là trung bình điều hoà của hai độ đo Precision và Recall.

(5) Độ đo Specitivity (Độ đặc hiệu)

- Định nghĩa: Specitivity = TN/(TN+FP).

- Ý nghĩa: Độ đo Specitivity đánh giá khả năng một dữ liệu là phần tử âm được bộ phân lớp cho ra kết quả chính xác.

1.2. Tổng quan về học máy

1.2.1. Khái niệm về học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể .

1.2.2 Phân loại các loại học máy

- Học có giám sát

- Học không giám sát
- Học bán giám sát

1.3. Thuật toán cây quyết định

1.3. Xây dựng Cây quyết định dựa trên Entropy

Khái niệm Entropy [5] của một tập S được định nghĩa trong lý thuyết thông tin là số lượng mong đợi các bit cần thiết để mã hóa thông tin một thành phần rút ra một cách ngẫu nhiên từ tập S về lớp của nó. Đối với trường hợp tối ưu, mã sẽ có độ dài ngắn nhất. Theo lý thuyết thông tin, một mã có độ dài tối ưu sẽ được gán $-\log_2 p$ bits cho một thông điệp có xác suất là p.

Độ đo Entropy của tập mẫu S được định nghĩa bởi công thức sau:

$$\text{Entropy(S)} = \sum_{i=1}^{C} - p_i \log_2 p_i$$

Về bản chất, độ đo Entropy sẽ phản ánh mức độ không đồng nhất của tập mẫu S. Entropy là một độ đo để đo độ pha trộn dữ liệu của một tập mẫu, Entropy càng nhỏ thì tập mẫu càng đồng nhất.

1.3.2. Xây dựng cây quyết định dựa trên Gini index

Công thức Gini index thường được sử dụng phổ biến hơn Goodness of Split, là phương pháp hướng đến đo lường tần suất một đối tượng dữ liệu ngẫu nhiên trong tập dữ liệu ban đầu được phân loại không chính xác, trên cơ sở đối tượng dữ liệu đã nằm trong một tập con đã được phân ra từ dữ liệu ban đầu, có dán nhãn để thể hiện thuộc tính chung bất kỳ của các đối tường còn lại trong tập con này, giá trị phân loại chính là nhãn của tập con.

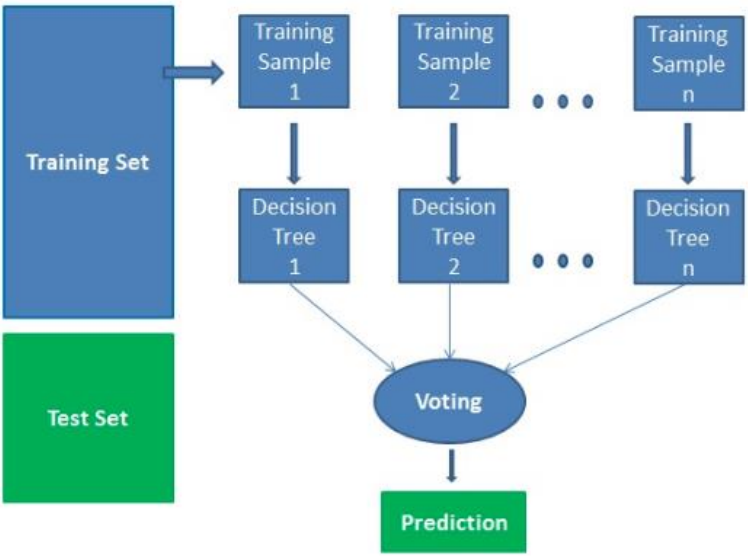
Gini index cũng chính là chỉ số đo lường mức độ đồng nhất hay mức độ nhiễu loạn của thông tin. Công thức Gini có thể áp dụng cho cả biến định tính và biến định lượng.

Gini index cho phép chúng ta đánh giá sự tối ưu của từng các phân nhánh thông qua xác định mức độ thuần khiết của từng node trong mô hình cây quyết định. Nếu tất cả các điểm dữ liệu nằm về cũng một lớp thì thể hiện sự đồng nhất không có nhiễu loạn ứng với Gini bằng 0, và sẽ

càng lớn nếu các điểm dữ liệu khác biệt nhau và lớn nhất bằng 1.

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

1.3.3. Thuật toán Rừng ngẫu nhiên



(Nguồn: Internet)

Hình 1.1: Thuật toán rừng ngẫu nhiên

Rừng ngẫu nhiên được đề xuất vào năm 2001 [2]. Đây là thuật toán phân loại có kiểm định dựa trên cây

quyết định và kỹ thuật Bagging and Bootstrapping đã được cải tiến. Bootstrapping là một phương pháp rất nổi tiếng trong thống kê được giới thiệu bởi Efron vào năm 1979. Phương pháp này được thực hiện như sau: từ một quần thể ban đầu lấy ra một mẫu L = (x_1,x_2,..,x_n) gồm n thành phần để tính toán các tham số mong muốn. Trong các bước tiếp theo lặp lại b lần tạo ra mẫu L_b cũng gồm n phần bằng cách lấy lại mẫu với sự thay thế các thành phần trong mẫu ban đầu sau đó tính toán các tham số mong muốn. Phương pháp Bagging được xem như là một phương pháp tổng hợp kết quả có được từ các bootstraping sau đó huấn luyện mô hình từ các mẫu ngẫu nhiên này và cuối cùng đưa ra dự đoán phân loại dựa vào số phiếu bầu cao nhất của lớp phân loại. Cây quyết định là một sơ đồ phát triển có cấu trúc dạng cây phân nhánh đi từ gốc cho đến lá, giá trị các lớp phân loại của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ gốc (tức là dữ liệu đầu vào) đến lá (tức là các kết quả phân loại dự đoán đầu ra), đường đi này biểu diễn sự phân lớp của mẫu đó. Mỗi sơ đồ cây

trong tập mẫu được tạo thành từ tập hợp các dữ liệu huấn luyện được lựa chọn ngẫu nhiên để huấn luyện mô hình phân loại Rừng ngẫu nhiên (mỗi tập mẫu bootstrap sẽ cho ra một cây và n cây tương ứng với n bootstrap). Khi một tập mẫu được rút ra từ tập huấn luyện (bootstrap) với sự thay thế có hoàn lại, thì thông thường có khoảng 1/3 các phần tử không nằm trong mẫu này và vì thế chúng không tham gia vào quá trình huấn. Điều này có nghĩa là chỉ có khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán để phân loại và 1/3 các phần tử này dùng để kiểm tra sai số. Dữ liệu kiểm tra được sử dụng để ước lượng sai số tạo ra từ việc kết hợp các kết quả phân loại riêng lẻ sau đó được tổng hợp trong mô hình Rừng ngẫu nhiên cũng như dùng để ước tính các biến quan trọng.

1.4. Thư viện Scikit-learn

Là một thư viện mạnh mẽ có thể mang các thuật toán học máy (machine learning) vào trong một hệ thống thích hợp nhất. Thư viện này tích hợp rất nhiều thuật toán hiện

đại và cố điển hỗ trợ việc học và tiến hành đưa ra các giải pháp hữu ích cho bài toán học máy một cách đơn giản.

1.5. Pycharm

1.5.1. Giới thiệu

Pycharm là một nền tảng kết kết hợp được JetBrains phát triển như một IDE (Môi trường phát triển tích hợp) để phát triển các ứng dụng cho lập trình trong Python. Một số ứng dụng lớn như Tweeter, Facebook, Amazon và Pinterest sử dụng Pycharm để làm IDE Python của họ. Bài viết dưới đây sẽ giới thiệu chi tiết cho bạn về Pycharm cũng như hướng dẫn cách cài đặt và sử dụng Pycharm

1.5.2. Các tính năng của Pycharm

Pycharm có thể chạy trên Windows, Linux, hoặc Mac OS. Ngoài ra, nó cũng chứa các Mô đun và các gói giúp các lập trình viên phát triển phần mềm bằng Python trong thời gian ngắn với ít công sức hơn. Hơn nữa, nó cũng có khả năng tùy chỉnh theo yêu cầu của nhà phát triển.

CHƯƠNG 2– PHƯƠNG PHÁP KHUYẾN NGHỊ GÓI CƯỚC

2.1. Phân tích các yếu tố ảnh hưởng tới gói cước phù hợp với khách hàng

Việc chọn gói cước phù hợp với khách hàng phụ thuộc vào nhiều yếu tố, trong phần này luận văn sẽ đi sâu phân tích các yếu tố ảnh hưởng trực tiếp đến việc lựa chọn gói cước phù hợp cho khách hàng.

2.1.1. Các yếu tố về khách hàng

Các yếu tố phi chất lương là các yếu tố được hình thành gồm:

Tên Thành phố, Quận, Huyện: Như chúng ta đã biết khách hàng tuy sử dụng cùng 1 loại hình dịch vụ tuy nhiên do tập quán sinh hoạt văn hóa…. 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: Doanh nghiệp, Tổ chức, Cá nhân…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.

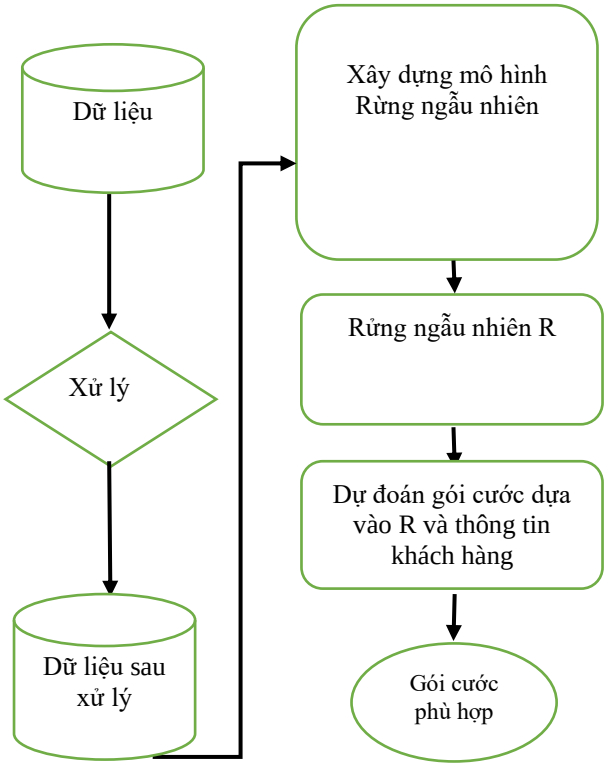
Độ tuổi khách hàng: Độ tuổi khách hàng phần nào đó thể hiện nhu cầu sử dụng dịch vụ của khách hàng. Ví dụ những người trẻ tuổi có nhu cầu sử dụng Internet tốc độ cao hơn để phục vụ cho các công việc online hoặc chơi game, xem phim trực tuyến. Những người lớn tuổi thì có xu hướng sử dụng dịch vụ MyTV để xem truyền hình, thời sự…

2.1.2. Các yếu tố về chất lượng dịch vụ

Tất cả mọi ngành nghề kinh doanh chất lượng sản phẩm dịch là linh hồn của doanh nghiệp, chất lượng càng cao thì sản phẩm được khách hàng ưu chuộng, doanh nghiệp bán được nhiều sản phẩm doanh thu mang về càng nhiều và cứ như thế doanh nghiệp ngày một phát triển, Viễn thông Tây Ninh cũng vậy, vì đã xác định chất lượng là mục tiêu hàng đầu để luôn cải thiện và hoàn chỉnh ngày một tốt hơn, từ đó có nhiều giải pháp để thực hiện, chất lượng gồm chất lượng của dịch vụ và chất lượng phục vụ.

2.2. Mô hình dự đoán gói cước cho khách hàng

Để tiến hành dự đoán gói cước phù hợp với khách hàng ta sử dụng mô hình được mô tả như trong Hình 2.4 như sau:



Hình 2.1: Mô hình thực nghiệm dự đoán

2.3. Sử dụng thuật toán phân lớp Rừng ngẫu nhiên thông qua bộ thư viện Scikit-learn

Để cài đặt scikit-learn trước tiên phải cài thư viện SciPy (Scientific Python). Những thành phần gồm:

- Numpy: Gói thư viện xử lý dãy số và ma trận nhiều chiều

- SciPy: Gói các hàm tính toán logic khoa học

- Matplotlib: Biểu diễn dữ liệu dưới dạng đồ thị 2 chiều, 3 chiều

- IPython: Notebook dùng để tương tác trực quan với Python

- SymPy: Gói thư viện các kí tự toán học

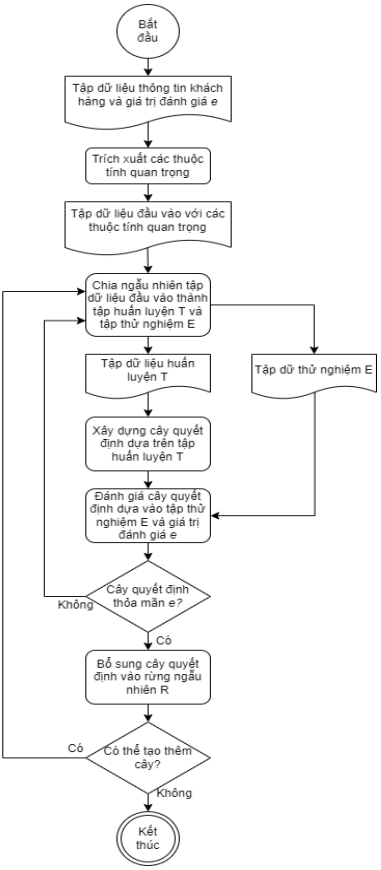
- Pandas: Xử lý, phân tích dữ liệu dưới dạng bảng

2.4. Sử dụng Pycharm để xây dựng ứng dụng web

Ứng dụng web được xây dựng bằng thư viện Flask trên ngôn ngữ Python. Server được xây dựng bằng ngôn ngữ Python để tiện cho việc truy xuất các model một cách dễ dàng hơn so với các ngôn ngữ khác.

CHƯƠNG 3 - XÂY DỰNG MÔ HÌNH

Quá trình để xây dựng rừng ngẫu nhiên được biểu diễn qua lưu đồ giải thuật như sau.



Hình 3.1: Lưu đồ giải thuật xây dựng rừng ngẫu nhiên

3.1. Dữ liệu

3.1.1. Thu thập dữ liệu

Hiện tại, các quy trình nghiệp vụ tại VNPT Tây Ninh đều được thao tác, thực hiện trên hệ thống thông tin Điều hành sản xuất kinh doanh (ĐHSXKD), đây là một hệ sinh thái lớn trong hệ thống quản lý của VNPT.

Hệ thống này cũng quản lý tất cả các việc thu thập thông tin khách hàng, quản lý thuê bao và các vấn đề liên quan. Vì vậy dữ liệu trong nghiên cứu này được trích xuất một phần từ cơ sở dữ liệu của hệ thống.

Dữ liệu thông tin khách hàng sau khi thu thập từ hệ thống ĐHSXKD cần thực hiện các bước tiền xử lý để loại bỏ các mẫu nhiễu trong tập dữ liệu như các dòng trống, các dòng không có giá trị.Các thông tin khách hàng từ tập dữ liệu sẽ được trích xuất để lấy các thuộc tính quan trọng với quá trình đề xuất gói cước, các thông tin được trích xuất cụ thể như sau:

Bảng 3.1 Bảng số trường và ý nghĩa từng trường dữ liệu

TT	Tên trường dữ liệu	Ý nghĩa	Kiểu dữ liệu
	TEN_DVDB	Địa chỉ lắp đặt thuê bao	Liệt kê
	TUOI	Độ tuổi khách hàng	Số
	TEN_LOAIKH	Loại khách hàng là cá nhân hay doanh nghiệp	Liệt kê
	TEN_NHOM	Tên nhóm khách hàng	Liệt kê
	IP_TINH	Nhu cầu sử dụng IP tĩnh	Liệt kê
	TOC_DO	Băng thông cần thiết cho khách hàng	Số
	NAM_DK	Năm khách hàng đăng ký dịch vụ	Số

	GOI	Gói cước khách hàng sử dụng	Liệt kê

3.1.2. Xử lý dữ liệu

Việc xử lý các dữ liệu ngoại lai sẽ giúp tăng cao độ chính xác cho các mô hình dự đoán hay các báo cáo doanh nghiệp một cách đáng kể. Từ kết quả của dữ liệu thu thập được, chúng ta có thể thấy rằng, dữ liệu vẫn một số trường hợp có thuộc tính có giá trị null khi không có thông tin.

3.1.3. Mã hóa dữ liệu

Mã hóa dữ liệu là quá trình bắt buộc sử dụng các phương pháp mã hóa cụ thể để tạo ra các giá trị phân nhóm cụ thể cho từng đặc trưng.

Từ kiểu dữ liệu có thể thấy, một số trường đang có kiểu dữ liệu chuỗi, vì vậy để có thể dễ dàng phân tích và đặc biệt là phục vụ cho các mô hình học máy ta sẽ chuyển chúng về dạng số. Phương pháp mã hóa dữ liệu sẽ dùng là phương pháp Label Encoder.

3.2. Xây dựng mô hình khuyến nghị gói cước dựa vào thuật toán rừng ngẫu nhiên

Một khách hàng với các thông tin quan trọng sau khi được trích xuất và chuẩn hóa được chuyển về dạng véc tơ để làm đầu vào xây dựng Rừng ngẫu nhiên với đầu vào và đầu ra như sau:

Đầu vào: tập dataset Y là tập dữ liệu thông tin khách hàng với các thuộc tính độ tuổi, loại khách hàng, tên nhóm khách hàng, địa bàn khách hàng đăng ký dịch vụ, nhu cầu sử dụng IP tĩnh, năm đăng ký bắt đầu sử dụng dịch vu, băng thông cần thiết, gói cước khách hàng sử dụng và giá trị ngưỡng đánh giá σ (giá trị trong luận văn sử dụng là giá trị accuracy score và điều kiện là $\sigma > 0.8$).

Đầu ra: Rừng ngẫu nhiên với tập hợp các cây quyết định tối ưu R.

3.2.1. Lấy mẫu dữ liệu cho việc xây dựng cây quyết định trong rừng ngẫu nhiên

Chia tập dataset Y ngẫu nhiên thành 2 phần: tập dữ liệu để kiểm tra E (20%) và tập dữ liệu để huấn luyện T (80%). Sử dụng T để xây dựng Rừng ngẫu nhiên. Dữ liệu huấn luyện T với các thuộc tính độ tuổi, loại khách hàng, tên nhóm khách hàng, địa bàn khách hàng đăng ký dịch

vụ, như cầu sử dụng IP tĩnh để xây dựng mô hình phân lớp. Tổng cộng có 5 thuộc tính dùng cho việc phân lớp dữ liệu.

3.2.2. Xây dựng cây quyết định trong rừng ngẫu nhiên

Bước 1: CART tách biến đầu tiên trong tất cả các biến tại các điểm phân tách có thể xảy ra, tại tất cả các giá trị mà biến giả định có trong tập mẫu. Tại mỗi điểm phân chia có thể có của một biến, mẫu phân tách thành hai nút con. Các trường hợp có câu trả lời "có" cho câu hỏi được đặt ra được gửi đến nút bên trái và những trường hợp phản hồi "không" được gửi đến đúng nút bên phải.

Bước 2: CART sau đó áp dụng tiêu chí phân tách dựa trên công thức GINIsplit tại mỗi điểm phân tách và đánh giá mức giảm tạp chất đạt được bằng cách sử dụng công thức:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Bước 3: CART chọn cách tách tốt nhất cho các biến với sự phân chia mà sự giảm tạp chất là cao nhất. Ba bước trên được lặp lại cho mỗi biến còn lại ở nút gốc.

Bước 4: CART sau đó xếp hạng tất cả các cách chia tốt nhất trên mỗi biến theo sự giảm tạp chất đạt được bằng mỗi lần tách và chọn biến cùng điểm tách mà nó làm giảm tạm chất của nút gốc và nút trung gian nhiều nhất.

Bước 5: CART sau đó gán lớp cho các nút này theo quy tắc giảm thiểu phân loại. CART có một thuật toán tích hợp để cho phép người dùng xác định điểm gán lớp trong quá trình tách. Mặc định là 1 đơn vị hoặc bằng giá trị phân loại. Bởi vì thủ tục CART là đệ quy, các bước 1 - 5 được áp dụng nhiều lần cho mỗi phần tử không phải nút lá ở mỗi giai đoạn kế tiếp.

CART dừng quá trình phân tách khi:

- Chỉ có một mẫu trong mỗi nút.

- Tất cả các mẫu trong mỗi nút con có phân loại giống hệt nhau của các biến phân lớp, tức là không thể tách.

- Cây đạt đủ độ sâu theo cài đặt (max_depth=10).

3.2.3. Xây dựng rừng ngẫu nhiên

Cây quyết định sau khi xây dựng xong nếu thỏa giá trị đánh giá ban đầu sẽ được cập nhật vào rừng ngẫu nhiên R. Việc đánh giá dựa trên tập thử nghiệm E và giá trị đánh giá σ. Độ chính xác của cây quyết định, cụ thể là giá trị F1 Score của cây sẽ được tính dựa vào kết quả phân loại tập thử nghiệm E, nếu Accuracy Score > 0.8 thì cây sẽ được đưa vào rừng ngẫu nhiên và ngược lại.

3.3. Xây dựng ứng dụng web

Sử dụng kết quả mô hình khuyến nghị gói cước dựa trên rừng ngẫu nhiên xây dựng được ở mục 3.2 để đưa vào ứng dụng web được xây dựng trên môi trường Pycharm.

Ứng dụng web được thiết kế để khách hàng hoặc nhân viên VNPT nhập các thông tin cần thiết là các thuộc tính phân lớp trong mô hình rừng ngẫu nhiên như: Tuổi khách hàng, Tốc độ mong muốn, Đơn vị cung cấp, Loại khách hàng, Nhóm khách hàng, nhu cầu dùng IP tĩnh.

CHƯƠNG 4 – PHÂN TÍCH VÀ ĐÁNH GIÁ

4.1. Phân tích độ chính xác của mô hình

Để đánh giá độ chính xác của mô hình đã xây dựng ta dựa vào kết quả phân lớp tập thử nghiệm là tập dữ liệu được trích xuất từ tập dataset ban đầu với 1000 mẫu thông tin khách hàng được xây dựng thủ công, trường thuộc tích gói cước được chọn từ kinh nghiệm của chuyên viên tư vấn bán hàng có kinh nghiệm để cho kết quả phù hợp với khách hàng nhất.

Để đánh giá mức độ ảnh hưởng của các tham số quan trọng trong quá trình xây dựng rừng ngẫu nhiên là max_depth (độ sâu của cây quyết định) và n_ estimators (số lượng cây quyết định trong rừng ngẫu nhiên), các thử nghiệm được thực hiện 10 lần và các kết quả thu được dựa trên giá trị trung bình của các lần chạy như bảng kết quả đánh giá sau.

Bảng 4.2: Giá trị Accuracy Score với hai tham số quan trọng của rừng ngẫu nhiên

max_depth\n_estimators	20	30	50	100	200
5	88.28	88.28	88.39	88.31	88.3
10	88.42	**88.51**	88.43	88.47	88.39
15	88.23	88.25	88.32	88.31	88.27
20	87.74	87.79	87.64	87.56	87.46
25	87.32	87.45	87.54	87.36	87.31

Dựa vào các chỉ số đánh giá độ chính xác của mô hình trên có thể đưa ra các nhận định sau:

- Mô hình cho độ chính xác cao nhất với giá trị max_depth = 10, giá trị n_estimators = 30 với các giá trị Accuracy Score = 88.51%.

- Số lượng thuộc tính của khách hàng còn ít nên chưa tạo được sự bao quát cho mô hình nên độ chính xác của mô hình còn chưa thực sự cao. Việc sử dụng ít thuộc tính thông tin khách hàng dẫn đến trường hợp có thể một thuộc tính thực sự chưa ảnh hưởng đến việc lựa chọn gói

cước cho khách hàng nhưng lại có giá trị ảnh hưởng cao trong mô hình khuyến nghị gói cước. Vì thế ta tiến hành đánh giá mức độ quan trọng của các thuộc tính trong mô hình để xác định các thuộc tính ảnh hưởng đến kết quả khuyến nghị gói cước có phải là yếu tố thực tế mang tính quyết định đến việc lựa chọn gói cước cho khách hàng hay không.

4.2. Xác định mức độ quan trọng của các thuộc tính

Mức độ quan trọng của các thuộc tính được xác định bằng độ giảm của chỉ số gini tại mỗi nút trong quá trình xây dựng cây quyết định. Độ giảm chỉ số gini càng nhiều ứng với mức độ quan trọng của thuộc tính càng cao.

CHƯƠNG 5 - KẾT LUẬN

5.1. Kết quả đạt được

5.1.1. Về mặt lý thuyết

Khai thác được tập dữ liệu thông tin khách hàng sử dụng Internet của VNPT Tây Ninh để xây dựng mô hình Khuyến nghị gói cước cho khách hàng.

Ứng dụng Trí tuệ nhân tạo (AI), Machine Learning, các thuật toán học máy vào việc khuyến nghị gói cước cho khách hàng.

Khai thác được các thuật toán phân lớp dữ liệu, cụ thể là mô hình cây quyết định và rừng ngẫu nhiên. Nắm bắt được quá trình xây dựng một cây quyết định dựa trên giá trị gini index hay entropy và quá trình xây dựng một rừng ngẫu nhiên dựa trên các cây quyết định.

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 rừng ngẫu nhiên xây dựng được.

5.1.2. Về mặt thực tiễn

Luận văn đã đưa ra được giải pháp khuyến nghị gói cước phù hợp với khách hàng sử dụng dịch vụ Internet của VNPT dựa vào việc phân tích tập dữ liệu khách hàng hiện có. Việc này sẽ là tiền đề để xây dựng một công cụ tư vấn bán hàng tự động thay thế cách tư vấn truyền thống nhân công mất thời gian và nhân lực nhưng đôi khi lại cho kết quả gói cước tư vấn cho khách hàng chưa thực sự phù hợp với khách hàng đối với các nhân viên chưa có kinh nghiệm.

Mô hình trên có thể hỗ trợ khách hàng chủ động tìm kiếm các gói cước phù hợp với bản thân khi cung cấp các thông tin cần thiết để chọn ra gói cước phù hợp nhất.

Xây dựng thành công mô hình khuyến nghị gói cước, phân tích và đánh giá mô hình xây dựng được để hiểu rõ hơn về các yếu tố ảnh hưởng đến việc lựa chọn một gói cước phù hợp cho khách hàng. Từ đó cũng rút ra được các điểm còn thiếu sót để có thể tiến hành thay đổi trong công tác thực hiện thực tế để mang lại trải nghiệm

tốt nhất cho khách hàng sử dụng dịch vụ Internet của VNPT.

5.2. Hạn chế

Kết quả khuyến nghị gói cước đạt được chỉ ở mức tốt chứ chưa thật sự cao. Kết quả đạt được chưa bao quát được hết các trường hợp. Dữ liệu thông tin khách hàng cần bổ sung thêm các thuộc tính thực tế tác động đến việc lựa chọn gói cước của khách hàng như số lượng thiết bị cần sử dụng Internet, cơ sở hạ tầng lắp đặt (nhà cấp 4, nhà lầu, nhà trọ...), các mục đích sử dụng chính khi lắp đặt Internet...

Mô hình rừng ngẫu nhiên trong luận văn còn ở mức cơ bản, chưa phân tích sâu vào các tham số để phù hợp với mô hình dữ liệu thông tin khách hàng sử dụng Internet của VNPT.

5.3. Hướng phát triển

Điều chỉnh công tác tư vấn bán hàng sang hướng tối ưu hóa trải nghiệm cho khách hàng hơn, thu thập thêm các thông tin mang tính cá nhân hóa với từng khách hàng như số lượng thiết bị cần sử dụng Internet, cơ sở hạ tầng

lắp đặt (nhà cấp 4, nhà lầu, nhà trọ...), các mục đích sử dụng chính khi lắp đặt Internet... để chọn ra gói cước phù hợp nhất. Sau đó sử dụng tập dữ liệu mới với các thuộc tính bổ sung để xây dựng lại một mô hình khuyến nghị gói cước phù hợp nhất với nhu cầu sử dụng của từng khách hàng. Từ đó gia tăng trải nghiệm của khách hàng khi sử dụng dịch vụ của VNPT.

Tiến hành áp dụng giúp hỗ trợ việc tư vấn bán hàng cho nhân viên kinh doanh mới, bán hàng online trên các website và nền tảng thông tin số khác của Viễn thông Tây Ninh. Với mục đích thực hiện tư vấn gói cước phù hợp, nhanh chóng và có độ tin cậy cao với từng khách hàng riêng biệt.