HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HOÀNG MẠNH HƯNG

ỨNG DỤNG MÁY HỌC ĐỂ DỰ ĐOÁN NGHỀ NGHIỆP CỦA THUÊ BAO DI DỘNG

CHUYÊN NGÀNH : **HỆ THỐNG THÔNG TIN**

MÃ SỐ: **8.48.01.04**

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – 2021

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học PGS. TS. Trần Quang Anh

Phản biện 1: PGS TS Nguyễn Hà Nam

Phản biện 2: PGS TS Ngô Quốc Tạo

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: 12 giờ 00 ngày 28 tháng 8 năm 2021

Có thể tìm hiểu luận văn tại:

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

MỤC LỤC

CHƯƠNG 1 – TỔNG QUAN NGHIÊN CỨU

1.1. Mô tả chung về nghiên cứu

1.1.1 Giới thiệu

Việc xác định được khách hàng là ai hiện là mối quan tâm hàng đầu của các nhà cung cấp sản phẩm và dịch vụ bán hàng. Nhờ xác định được chân dung khách hàng mà các doanh nghiệp có thể đạt được hiệu quả tối đa khi thực hiện các chiến dịch quảng bá sản phẩm, nhắm vào những đối tượng cụ thể và có các cách tiếp cận hợp lý nhất. Có rất nhiều biện pháp để xác định rõ ràng một bức tranh hoàn chỉnh về chân dung khách hàng, nhưng trong nghiên cứu này chúng tôi sẽ tập trung vào việc xác định chân dung khách hàng từ tập thuê bao viễn thông. Đóng góp chính trong công việc của chúng tôi là phát triển một mô hình dự đoán nghề nghiệp của thuê bao di động, giúp các nhà khai thác viễn thông dự đoán được đối tượng khách hàng của mình, từ đó có thể đưa ra các gói sản phẩm phù hợp cũng như cung cấp các dịch vụ giá trị gia tăng khác để thu hút cũng như giữ chân khách hàng, gia tăng lợi nhuận doanh nghiệp.

Mô hình được phát triển trong nghiên cứu này sử dụng các kĩ thuật học máy cho mục đích phân lớp nhị phân dựa trên bộ dữ liệu là các đặc trưng được xây dựng từ toàn bộ các bản ghi chi tiết về cuộc gọi, sử dụng dịch vụ để dự đoán một thuê bao có là sinh viên hay không. Để đo lường hiệu suất của mô hình, thước đo tiêu chuẩn AUC được sử dụng và giá trị AUC đạt được là 94,6% dựa trên thuật toán XGBoost. Mô hình được chuẩn bị và thử nghiệm thông qua ứng dụng Spark và H20 và làm việc trên bộ dữ liệu lớn được cung cấp và mã hóa từ một trong các công ty viễn thông hàng đầu tại Việt Nam. Bộ dữ liệu chứa tất cả các thông tin CDR của khách hàng đã được mã hóa số thuê bao để đảm bảo tính bảo mật và ATTT của khách hàng được sử dụng để huấn luyện, thử nghiệm và đánh giá mô hình. Nghiên cứu cũng thử nghiệm 4 thuật toán: Rừng ngẫu nhiên, mô hình tuyến tính tổng quát (GLM), máy tăng cường Gradien “GBM” và tăng cường độ dốc cao “XGBoost”. Tuy nhiên kết quả tốt nhất đã thu được bằng cách sử dụng thuật toán XGBoost, và thuật toán này sẽ được sử dụng để phân loại nghề nghiệp trong bài nghiên cứu này.

1.1.2 Đặt vấn đề và giải quyết bài toán

Trong bài nghiên cứu này, chúng tôi sẽ tập trung vào một phần nhỏ trong bức tranh tổng quát về chân dung khách hàng là nghề nghiệp của thuê bao di động, cụ thể là đánh giá xem thuê bao đó có phải là sinh viên hay không. Chúng tôi chuyển về bài toán phân loại nhị phân để đơn giản hóa quá trình lựa chọn và đánh giá mô hình. Dữ liệu được sử dụng bao gồm tất cả các CDR của thuê bao trong suốt 6 tháng trước thời điểm lấy mẫu. Chúng tôi sẽ sử dụng tập dữ liệu này để tổng hợp và trích xuất các đặc trưng cho từng khách hàng, sau đó sử dụng các mô hình học máy để dự đoán xem một thuê bao có là sinh viên hay không.

1.2. Các đặc điểm về dữ liệu nhà mạng

1.2.1 Giới thiệu về dữ liệu của nhà mạng

Là một trong 3 công ty viễn thông lớn nhất Việt Nam [6], đối tác của chúng tôi có rất nhiều loại dữ liệu bao gồm cả dữ liệu sử dụng dịch vụ của người dùng lẫn dữ liệu hoạt động của hệ thống. Các loại này được phân loại như sau :

- Dữ liệu khách hàng: Nó chứa tất cả các thông tin liên quan đến dịch vụ và hợp đồng của khách hàng.

- Dữ liệu về vị trí: Thông tin về vị trí xảy ra các sự kiện của người dùng cũng được lưu lại dưới dạng mã vị trí.

- Dữ liệu về khiếu nại, chăm sóc khách hàng: Bao gồm các thông tin phản ánh dịch vụ từ khách hàng, cũng như các khiếu nại liên quan tới cước, gói dịch vụ mà khách sở hữu, quan tâm.

- Dữ liệu nhật ký mạng: Chứa các thông tin về tình trạng hoạt động của hệ thống, các bản ghi lưu lại lịch sử hoạt động của ứng dụng, log chi tiết về hệ thống, cũng như lịch sử thay đổi của các phiên bản nâng cấp hệ thống.

- Dữ liệu chi tiết cuộc gọi: Chứa các thông tin chi tiết về cuộc gọi, sms, mms, truy cập internet, vasp.. Dữ liệu này được tạo ra dưới dạng văn bản

- Thông tin thiết bị di động : Nó chứa thông tin về thương hiệu, kiểu máy, loại điện thoại di động, dùng 1 sim hay 2 sim, hỗ trợ 4g hay không..

Do vấn đề bảo mật chúng tôi chỉ có thể tiếp cận một số loại dữ liệu như dữ liệu về vị trí, dữ liệu về thiết bị di động và dữ liệu về chi tiết cuộc gọi.

1.2.2 Khối lượng dữ liệu lớn và phức tạp

Vì chúng tôi không biết thông tin nào có thể hữu ích cho quá trình xây dựng mô hình, vì thế chúng tôi phải xử lý tất cả các dữ liệu phản ánh hành vi và hoạt động của tất cả các khách hàng. Bộ dữ liệu chúng tôi sử dụng được lấy trong vòng 6 tháng đến thời điểm lấy mẫu, với trung bình mỗi ngày xấp xỉ 300GB dạng text, tương đương với việc chúng tôi phải xử lý tất cả hơn 50TB dữ liệu thô.

1.2.3 Dữ liệu đa dạng và trùng lặp

Dữ liệu CDR được đến từ nhiều nguồn khác nhau, do cách lấy, cấu trúc, cũng như cách lưu log khác nhau từ nguồn cung cấp dữ liệu. Do đó, dữ liệu bao gồm rất nhiều thông tin trùng lặp và phân tán của cùng một loại dữ liệu. Vì vậy, chúng tôi phải xử lý tất các các nguồn dữ liệu, hiểu nó và sau đó so sánh và chọn một hoặc kết hợp nhiều nguồn dữ liệu về một bảng thống nhất. Kết quả đạt được là đã rút gọn đi được 1 nửa số bảng và trường dữ liệu trùng lặp và không cần thiết.

1.2.4 Tập dữ liệu không cân bằng

Tập dữ liệu được tạo không cân bằng vì nó là một trường hợp đặc biệt của bài toán phân loại trong đó sự phân bố của một lớp thường không đồng nhất với một lớp khác. Tập dữ liệu là không cân bằng nếu một trong các danh mục của nó nhỏ hơn hoặc bằng 10% so với tập còn lại.

1.2.5 Giá trị bị mất

Các khách hàng khác nhau có thể có các gói dịch vụ khác nhau. Vì thế, có thể có khách hàng có những gói cước, dịch vụ hoặc sản phẩm mà khách hàng khác không có, trong khi họ lại có một số thứ khác.

Ngoài việc giá trị bị mất, còn có thể xảy ra trường hợp dữ liệu bị mất. Đó có thể do lỗi hệ thống, mất log, thiếu log hoặc lỗi xử lý sai dữ liệu, lỗi đường truyền khiến một số bản ghi bị mất mà không thể khôi phục lại.

1.2.6 Giá trị cố định

Sau khi khai phá dữ liệu, chúng tôi nhận thấy rằng khoảng 50% biến số chứa một hoặc hai giá trị rời rạc và khoảng 80% tất cả các biến phân loại có ít hơn 10 danh mục, 15% biến số và biến phân loại chỉ có một giá trị. Có những biến mà hầu hết giá trị của chúng là 0 hoặc là hằng số. Chúng tôi thấy rằng có khoảng 77% các biến số có hơn 97% giá trị của chúng là 0, là hằng số hoặc rỗng. Những kết quả này chỉ ra rằng, một lượng lớn các biến có thể loại bỏ do chúng không có giá trị hoặc giá trị cố định.

1.3. Phân nhóm nghề nghiệp và dữ liệu mẫu

1.3.1 Lý thuyết chọn mẫu

Tổng thể là tập hợp tất cả các đối tượng khảo sát. Mẫu là một tập hợp nhỏ những phần tử được lấy ra từ một tổng thể lớn, người ta sẽ nghiên cứu những mẫu đó để tìm ra đặc trưng của mẫu. Các đặc trưng của mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể do nó đại diện.Có hai cách chọn mẫu : Chọn mẫu theo xác suất và chọn mẫu phi xác suất

1.3.2 Mẫu nghiên cứu

Việc thu thập được số mẫu đại diện cho tất cả là rất khó khăn, chúng tôi đã sử dụng phương pháp chọn mẫu phi xác suất – chọn mẫu thuận tiện để lựa chọn mẫu. Cụ thể, chúng tôi lựa chọn ra 7,388 sinh viên từ các trường Đại học Hà Nội, Đại học Y Hà Nội và Đại học Đà Nẵng. Sau đó chúng tôi lựa chọn thêm hơn 8000 mẫu đối lập để tạo thành hơn 15000 nhãn. Nhưng sau khi thực hiện lấy đặc trưng, chúng tôi chỉ còn 6438 mẫu sinh viên và 6990 mẫu đối lập cho tổng thể 13428 mẫu. Cách lấy mẫu này có ưu điểm là chọn mẫu một cách thuận tiện, dễ tiếp cận và lấy thông tin. Nhưng có nhược điểm là không xác định được sai số lấy mẫu và không kết luận được tổng thể từ kết quả mẫu. Chúng tôi sẽ phải cải thiện bằng cách thu thập thêm nhiều mẫu ngẫu nhiên hơn, loại bỏ các đặc trưng có thể bị phân lập vào nhãn để đảm bảo mô hình có thể đạt được kết quả tốt nhất.

1.4. Kết luận

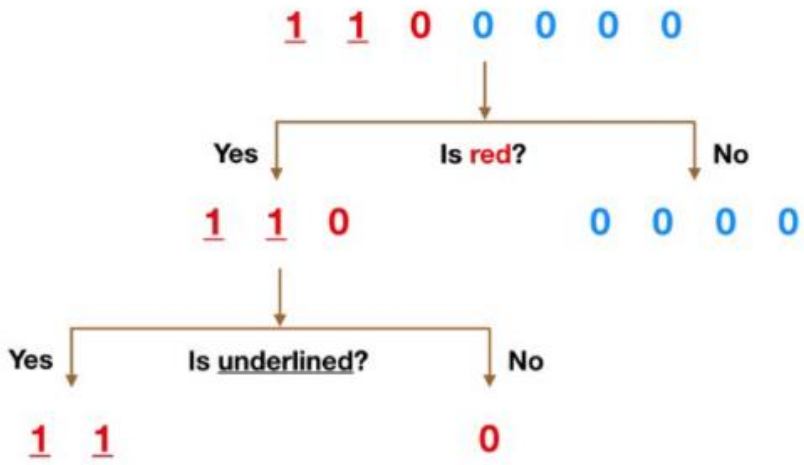
Như vậy, dữ liệu của nhà mạng là rất phức tạp và đồ sộ. Việc chuẩn hóa dữ liệu này hết sức mất thời gian và công sức. Để có thể tổng hợp ra những đặc trưng mạnh mẽ có tính phân loại cao, tôi phải thử đi thử lại nhiều thuật toán với các tham số khác nhau để đạt được một mô hình tốt nhất. Bằng việc sử dụng phần mềm hỗ trợ H2O, tôi có thể đẩy nhanh quá trình huấn luyện và kiểm tra. Chế độ mạnh mẽ nhất của H2O là AutoML, chế độ này sẽ sử dụng 4 mô hình cơ bản là eXtreme Gradient Boosting(XGBoost), Gradient Boosting Machine (GBM), General Linear Model (GLM), Distributed Random Forest (DRF) để thực hiện huấn luyện. Tôi sẽ giới thiệu mô hình lý thuyết và cách thức thực hiện huấn luyện và kiểm tra mô hình ứng với các thuật toán này trong các chương sau.

CHƯƠNG 2 – MỘT SỐ THUẬT TOÁN HỌC MÁY LIÊN QUAN

Trong chương này chúng ta sẽ tiếp cận một số thuật toán về học máy, từ các thuật toán đơn giản như cây quyết định đến thuật toán phức tạp hơn như XGBoost.

2.1 Cây quyết định

Cây quyết định (Decision Tree) là một mô hình thuộc nhóm thuật toán Học có giám sát (Supervised Learning). Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal. Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.



Hình 2.1 Mô hình diễu diễn cây quyết định

2.2 Rừng ngẫu nhiên

Rừng ngẫu nhiên là phương pháp xây dựng một tập hợp rất nhiều cây quyết định và sử dụng phương pháp bầu chọn để đưa ra quyết định về biến mục tiêu cần được dự báo. Random Forest (RF) là một tập hợp của hàng trăm cây quyết định (DF), trong đó mỗi cây quyết định được tạo nên ngẫu nhiên từ việc tái chọn mẫu

(chọn ngẫu nhiên một phần của dữ liệu để xây dựng) và ngẫu nhiên các biến từ toàn bộ các biến trong trong dữ liệu.

Với một cơ chế như vậy, Random Forest cho ta một kết quả chính xác rất cao nhưng đánh đổi bằng việc ta không thể hiểu cơ chế hoạt động của thuật toán này do cấu trúc quá phức tạp của mô hình này - do vậy thuật toán này là một trong những phương thức Black Box - tức ta sẽ bỏ tay vào bên trong và rút ra được kết quả chứ không thể giải thích được cơ chế hoạt động của mô hình.

2.3 Mô hình tuyến tính tổng quát

Mô hình tuyến tính tổng quát (GLM) là một sự tổng quát hóa linh hoạt của hồi quy tuyến tính thông thường cho phép các biến đáp ứng có mô hình phân phối lỗi khác với phân phối chuẩn . GLM tổng quát hóa hồi quy tuyến tính bằng cách cho phép mô hình tuyến tính có liên quan đến biến phản hồi thông qua một hàm liên kết và bằng cách cho phép độ lớn của phương sai của mỗi phép đo là một hàm của giá trị dự đoán của nó.

2.3 Các thuật toán boosting

Có hai thuật toán boosting được giới thiệu trong chương này là Extreme Gradient Boosting (XGBoost) và Gradient Boosting Machine (GBM). Cả hai thuật toán đều phát triển từ thuật toán máy tăng cường độ dốc (gradient boosting). Tuy nhiên XGBoost thường cho kết quả tốt hơn, do nó sử dụng phương thức chính thức hóa mô hình một cách chính quy hơn để kiểm soát việc quá vừa dữ liệu (over-fitting). XGBoost là thuật toán state-of-the-art nhằm giải quyết bài toán học có giám sát (supervised learning) cho độ chính xác khá cao. XGBoost nhận đầu vào là bảng dữ liệu với mọi kích thước và dạng dữ liệu bao gồm cả categorical mà dạng dữ liệu này thường được sử dụng nhiều hơn trong thực tế. Bên cạnh đó, XGboost có tốc độ huấn luyện nhanh, có khả năng scale để tính toán song song trên nhiều server, có thể tăng tốc bằng cách sử dụng GPU, nhờ vậy mà Big Data không phải là vấn đề của mô hình này.

XGBoost và GBM đều dựa trên cùng ý tưởng đó là boosting thông qua gradient descent trong không gian hàm số. Tuy nhiên, điều làm nên hiệu suất ấn tượng và khả năng tính toán của XGBoost nằm ở ba yếu tố:

- Engineering để tránh quá vừa dữ liệu (overfiting) như: lấy mẫu phụ theo hàng, theo cột, và cột trên mỗi cấp độ phân chia, áp dụng tăng cường chính quy với cả L1 và L2.
- Khả năng tận dụng tài nguyên hệ thống: tính toán song song trên CPU/GPU, tính toán phân tán trên nhiều server, tính toán khi tài nguyên bị giới hạn, tối ưu bộ nhớ đệm để tăng tốc huấn luyện.
- Và cuối cùng là khả năng xử lý các giá trị dữ liệu bị thiếu, tiếp tục huấn luyện bằng mô hình đã được xây dựng trước đó để tiết kiệm thời gian.

2.4 Đánh giá mô hình

2.4.1 Độ đo dùng trong phân loại

Khi xây dựng một mô hình Machine Learning, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình. Trong trường hợp phân loại hai lớp, chúng ta sử dụng ma trận nhầm lẫn để tính các độ đo. Giả sử ta có các chỉ số TP, FP, TN, FN lần lượt là dương tính đúng, dương tính sai, âm tính đúng, âm tính sai thì các độ đo được tính bởi công thức:

- Độ chính xác : $\text{accuracy} = \frac{TP+TN}{N}$
- Tỉ lệ dương đúng: $tpr = \frac{TP}{TP+FN}$
- Tỉ lệ dương sai: $fpr = \frac{FP}{TP+FN}$

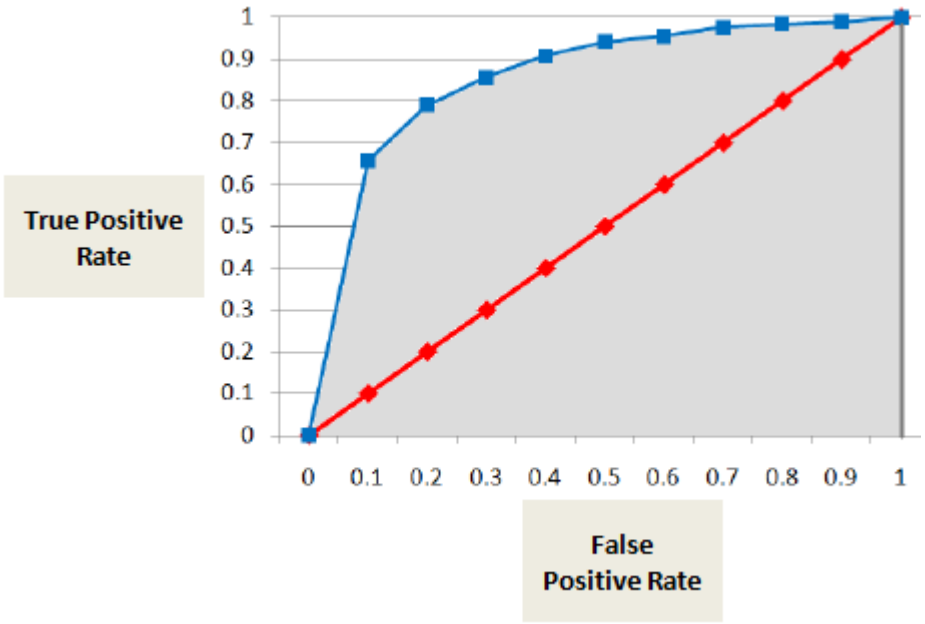
Các độ đo nói trên đều có giá trị nằm trong khoảng [0, 1]. Chúng ta sử dụng tpr và fpr để vẽ các đường cong ROC và tính AUC. Còn độ đo accuracy được dùng khi ta chỉ quan tâm tới độ chính xác nói chung.

2.4.2 ROC và AUC

Đường cong ROC (Receiver operating characteristic) và AUC (Area under Curve) được sử dụng để ước lượng và tính toán hiệu năng của mô hình phân loại. Nó đo lường khả năng phân biệt (discrimination power) của mô hình phân loại. Nói

một cách đơn giản, nó kiểm tra khả năng phân biệt các sự kiện trong mô hình phân lớp.

Đường cong ROC biểu diễn tỷ lệ dương tính đúng (tpr) so với tỉ lệ dương tính sai (fpr). AUC được tính là diện tính phía dưới đường cong ROC. AUC cung cấp một thước đo tổng hợp về hiệu suất trên tất cả các ngưỡng phân loại có thể có. Như hình bên dưới, AUC được tính là phần diện tích màu xám. Trong nhiều mô hình học máy, AUC được sử dụng làm thước đo chính để đánh giá mô hình.



Hình 2.4. Đường cong ROC và AUC

2.4.3 Đánh giá mô hình bằng kiểm tra chéo

Khi sử dụng các thuật toán phân loại và hồi quy, một công đoạn quan trọng là đánh giá độ chính xác của mô hình. Có nhiều cách để đánh giá mô hình, nhưng chúng tôi sử dụng kiểm tra chéo (cross-validation) để đạt được hiệu quả tốt nhất. Có 2 phương pháp kiểm tra chéo thường được sử dụng.

- Kiểm tra chéo với tập tách riêng: nghĩa là tách bộ huấn luyện làm hai tập riêng biệt, thường với tỉ lệ 70/30. Sau đó huấn luyện trên tập huấn luyện và kiểm tra trên tập kiểm thử.

- Kiểm tra chéo k-fold : phương pháp này chia tập mẫu thành k tập con, và lần lượt thực hiện k lần phương án lần lượt chọn 1 tập con làm tập kiểm thử và gộp k-1 tập còn lại làm tập huấn luyện.

2.5 Kết luận

Trong chương này tôi đã giới thiệu các thuật toán từ cơ bản đến nâng cao sẽ được áp dụng để huấn luyện mô hình. Đồng thời tôi cũng đưa ra các chỉ số để đánh giá chất lượng mô hình, từ đó tôi có thể quyết định lựa chọn sử dụng mô hình nào cho bước cuối cùng là huấn luyện và dự đoán toàn tập thuê bao. Khi đã chuẩn bị được dữ liệu, lựa chọn được các thuật toán sẽ sử dụng, tôi bắt đầu tiến vào bước cuối cùng, bước thực nghiệm. Từ dữ liệu đã có, tôi phải trích xuất ra các đặc trưng, sau đó lựa chọn các đặc trưng mạnh mẽ, cuối cùng áp dụng các thuật toán đã giới thiệu trong chương này để tiến hành huấn luyện, kiểm tra và rút ra kết luận cuối cùng.

CHƯƠNG 3 - ỨNG DỤNG HỌC MÁY ĐỂ PHÂN NHÓM NGHỀ NGHIỆP

3.1 Mô hình đề xuất

Có rất nhiều thuật toán có thể được sử dụng cho bài toán phân loại, nhất là bài toán phân loại nhị phân. Để có thể đánh giá xem thuật toán nào đạt chất lượng tốt nhất, chúng tôi sử dụng phần mềm H2O trên nền tảng của R. H2O cho phép hệ thống tự động lựa chọn các thuật toán phổ biến nhất trong danh sách bao gồm XGBoost, GBM, GLM, DRF (bao gồm DF và XFT), chạy lặp đi lặp lại nhiều lần và sắp xếp chúng theo độ chính xác AUC giảm dần. Sử dụng H2O có thể giúp chúng tôi đẩy nhanh quá trình training và testing, từ đó có thời gian tập trung vào việc xử lý dữ liệu và xây dựng đặc trưng phù hợp, nâng cao chất lượng mô hình.

3.2 Xử lý dữ liệu

3.2.1 Các bảng dữ liệu chính

Để chuẩn bị dữ liệu đầu vào cho bài toán, chúng tôi phải xử lý và làm sạch toàn bộ các dữ liệu dạng log giao dịch của nhà mạng. Các dữ liệu này được tổ chức thành các bảng riêng biệt và có cấu trúc, bao gồm các bảng cơ bản sau:

- Bảng log thoại và tin nhắn : Đây là các dữ liệu phát sinh do hành động gọi thoại, nhắn tin sms, mms của thuê bao bao gồm chi tiết về hướng cuộc gọi, thời gian, vị trí, thiết bị, tài khoản sử dụng..

- Bảng thông tin số dư hàng ngày: Đây là bảng tổng hợp số dư của tất cả các tài khoản của thuê bao tại thời điểm xuất báo cáo.

- Bảng thông tin cước : Đây là bảng tổng hợp hàng tháng, thể hiện chi tiết cước của các loại dịch vụ mà thuê bao sử dụng.

- Bảng nạp, ứng tiền : Thể hiện chi tiết các giao dịch nạp và ứng tiền, trả tiền của thuê bao.

- Bảng sử dụng dịch vụ mạng: Thể hiện thông tin chi tiết về hành vi sử dụng internet như dung lượng sử dụng, thời gian sử dụng, vị trí, thiết bị sử dụng, cước sử dụng …

- Bảng dịch vụ VAS: Đây là bảng thể hiện lịch sử đăng kí, hủy đăng kí hoặc gia hạn các dịch vụ giá trị gia tăng của khách hàng.

3.2.2 Xây dựng đặc trưng

Sau khi khảo sát và tổ chức lại dữ liệu, chúng tôi bắt đầu xây dựng các đặc trưng cho bài toán. Ban đầu, các đặc trưng được thiết kế ở mức tối đa nhất có thể bằng cách gom nhóm các thông tin trong một trường dữ liệu, tổ hợp các trường dữ liệu với nhau, tổ hợp các thông tin giống nhau trong các bảng lại với nhau và thực hiện các phép toán thống kê như lấy giá trị nhỏ nhất (min), lấy giá trị lớn nhất (max), lấy trung bình (avg), lấy tỉ lệ (ratio).. Ngoài ra, chúng tôi còn tổng hợp tất cả các tổ hợp tạo được theo từng mức thời gian, như theo từng khung giờ, theo ngày, tuần, tháng. Kết quả, chúng tôi đã xây dựng được một bộ hơn 5000 đặc trưng để phục vụ cho bước tiếp theo của bài toán.

3.2.2 Trích chọn đặc trưng

Trong bước trước, chúng tôi đã tập hợp được hơn 5000 đặc trưng tất cả. Tuy nhiên trong học máy, không phải cứ càng nhiều đặc trưng thì độ chính xác càng cao, mà ngược lại nó còn gây nhiễu và khiến quá trình học máy tốn thời gian hơn và thiếu chính xác hơn. Vì vậy, chúng tôi đã phải rút gọn bớt các đặc trưng mà vẫn đảm bảo được tính hiệu quả của mô hình.

Chúng tôi sử dụng phương pháp trọng số dấu hiệu (WOE - weight of evidence) để trích chọn các đặc trưng. Phương pháp này sẽ xếp hạng các đặc trưng thành mạnh, trung bình, yếu, không tác động,… dựa trên khả năng, sức mạnh dự đoán. Tiêu chuẩn xếp hạng sẽ là chỉ số giá trị thông tin IV (information value) được tính toán từ phương pháp WOE. Đồng thời mô hình cũng tạo ra các giá trị features cho mỗi biến. Giá trị này sẽ đo lường sự khác biệt trong phân phối giữa good và bad

Bằng cách này chúng tôi sẽ rút gọn và lấy ra được các đặc trưng mạnh mẽ nhất để xây dựng mô hình. Thực hiện công việc với mỗi bảng và lấy ra top 100 đặc trưng tốt nhất, chúng tôi rút gọn còn 811 đặc trưng để phục vụ cho giai đoạn tiếp theo.

Bảng 3.9 Bảng mô tả đặc trưng

isdn_key <int>	label <chr>	sim_age_months <int>	sim_age_days <int>	balance_weekday_ge_010k_dates <int>	balance_dates <int>
140210475	N	127	3865	17	92
139428761	Y	56	1709	37	92
175807197	N	NA	NA	-1	-1
160554310	N	61	1859	44	92
152913189	Y	45	1374	49	92
116201750	Y	66	2036	7	92
124637675	Y	68	2087	0	92
130703688	Y	96	2950	61	92
136241880	N	133	4064	52	92
122153068	Y	63	1920	0	92

1-10 of 12,102 rows | 1-6 of 813 columns

Previous 1 2 3 4 5 6 ... 100 Next

3.3 Thực nghiệm và kết quả

Sau khi xử lý dữ liệu và trích chọn xong các đặc trưng, chúng tôi tiến hành bước cuối cùng là huấn luyện và đánh giá mô hình.Để thực hiện nhanh việc huấn luyện và kiểm tra mô hình, chúng tôi sử dụng phần mềm H2O trên nền tảng R. Đây là những công cụ mạnh mẽ giúp các nhà phát triển dễ dàng thử nghiệm các mô hình một cách đơn giản và nhanh chóng.

AutoML của H2O có thể được sử dụng để tự động hóa các quy trình học máy, bao gồm đào tạo tự động và điều chỉnh nhiều mô hình trong giới hạn thời gian do người dùng chỉ định. Các thuật toán sẽ được sử dụng bao gồm ba mô hình XGBoost, GBM, GLM, DRM, DRF, XFT.. Tùy vào thời gian thiết lập cho phép mà AutoML sẽ chạy được số các thuật toán khác nhau, sau đó nó sẽ xếp hạng chúng theo tiêu chí tốt nhất ở trên đầu bảng.

Chúng tôi chia dữ liệu thành hai nhóm: nhóm đào tạo và nhóm thử nghiệm. Nhóm đào tạo gồm 90% tập dữ liệu nhằm mục đích đào tạo các thuật toán, nhóm kiểm tra chứa 10% tập dữ liệu sử dụng để kiểm tra các thuật toán. Cụ thể, dữ liệu thực tế của chúng tôi bao gồm 12102 bản ghi dành cho việc huấn luyện và 1286 bản ghi dành cho việc kiểm thử. Các tham số của thuật toán được tối ưu hóa bằng cách sử dụng xác thực chéo K-lần (K=9). Chúng tôi sử dụng R để thực thi H20. Dữ liệu để huấn luyện của chúng tôi bao gồm 811 đặc trưng.

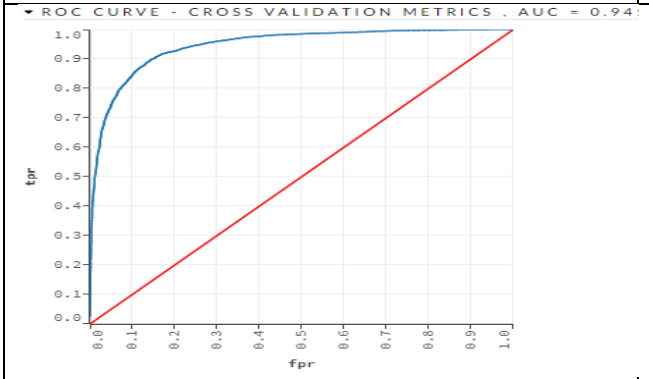
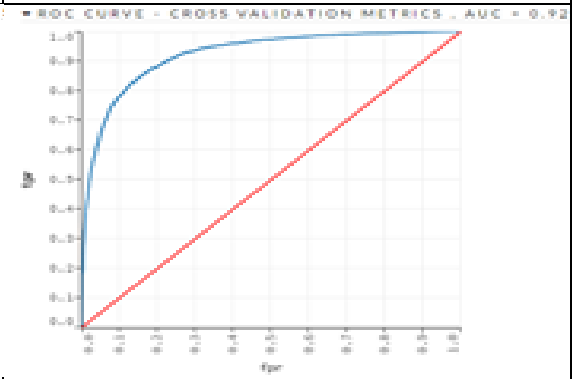
Kết quả sau khi cho H2O chạy AutoML như Bảng 3.10.

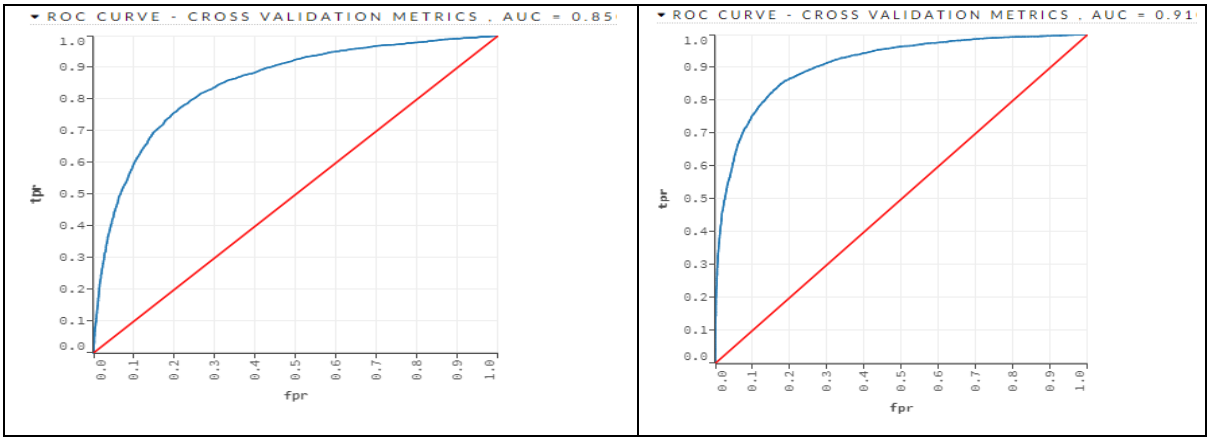
Bảng 3.10 Danh sách các thuật toán triển khai

	model_id	auc	logloss	aucpr	mean_per_class_
0	StackedEnsemble_AllModels_AutoML_20210609_002811	0.9497677153738137	0.28606582052759394	0.9469150884401346	0.1170641982526
1	StackedEnsemble_BestOfFamily_AutoML_20210609_002811	0.9478811899311	0.2914783971570033	0.9439908759675965	0.1212370710144
2	XGBoost_grid__1_AutoML_20210609_002811_model_2	0.9457839076817162	0.2976819808622501	0.9418718025765045	0.1233685027237
3	GBM_grid__1_AutoML_20210609_002811_model_1	0.9442256678268377	0.30215738936474423	0.940926021236419	0.1249659324988
4	XGBoost_grid__1_AutoML_20210609_002811_model_3	0.9435203995213612	0.30302015911450586	0.9396156570251059	0.1273919511089
5	XGBoost_grid__1_AutoML_20210609_002811_model_5	0.9426119282697425	0.3066418306252005	0.9389862440830209	0.1275566744436
6	GBM_grid__1_AutoML_20210609_002811_model_4	0.9424627907485016	0.30877330100129674	0.9390695254959427	0.1283730377720
7	XGBoost_grid__1_AutoML_20210609_002811_model_1	0.9418519252099509	0.31972009333174195	0.9381840493688117	0.1296144799173
8	GBM_5_AutoML_20210609_002811	0.9418080226643136	0.3111425804265272	0.9391739941025811	0.1305564859386
9	GBM_1_AutoML_20210609_002811	0.9417687371256536	0.3080610315331768	0.9388137867177171	0.1298012091640
10	GBM_grid__1_AutoML_20210609_002811_model_2	0.9415701648462729	0.3090118191897934	0.9392709830171573	0.1301226512001

Có thể nhận thấy, ngoài mô hình StackedEnsemble là mô hình tổng hợp các họ đặc trưng tốt nhất dựa trên H2O thì các mô hình có kết quả tốt nhất là sử dụng thuật toán GBM và XGBoost. Trong quá trình huấn luyện, mô hình XGBoost đang cho kết quả tốt nhất với giá trị AUC đạt 94.6%, GBM đạt giá trị AUC tốt nhất là 94,4%. Tôi sẽ chọn mô hình có kết quả tốt nhất của 4 thuật toán XGBoost, GBM, GLM và DRF để tiến hành xem xét và đánh giá chi tiết.

Trước hết chúng ta hãy xem xét đường cong ROC validation sau khi thực hiện xác thực 10-fold, có thể nhận thấy cả ba mô hình đều có khả năng phân loại rất tốt. Trong đó, ở quá trình kiểm thử, XGBoost là tốt nhất với mức AUC = 94.5%, GBM đạt 92,3%, DRF đạt 91,5%, còn GLM thì tệ nhất nhưng vẫn đạt 89,4%.

XGBoost	GBM
	
GLM	DRF



Hình 3.2 Đường cong ROC validation

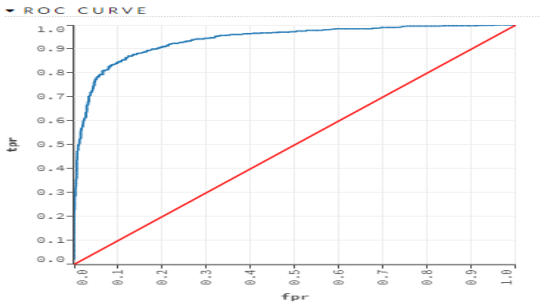
Như vậy ta có thể thấy, thuật toán XGBoost đang đạt hiệu quả cao nhất. Sử dụng mô hình này chúng ta kiểm tra lại kết quả trên tập dùng để thử nghiệm đã được tách ra từ trước cho kết quả như trên Bảng 3.11.

Bảng 3.11 Ma trận nhầm lẫn khi thực hiện dự đoán trên mẫu kiểm thử

PREDICTION - CONFUSION MATRIX ROW LABELS:

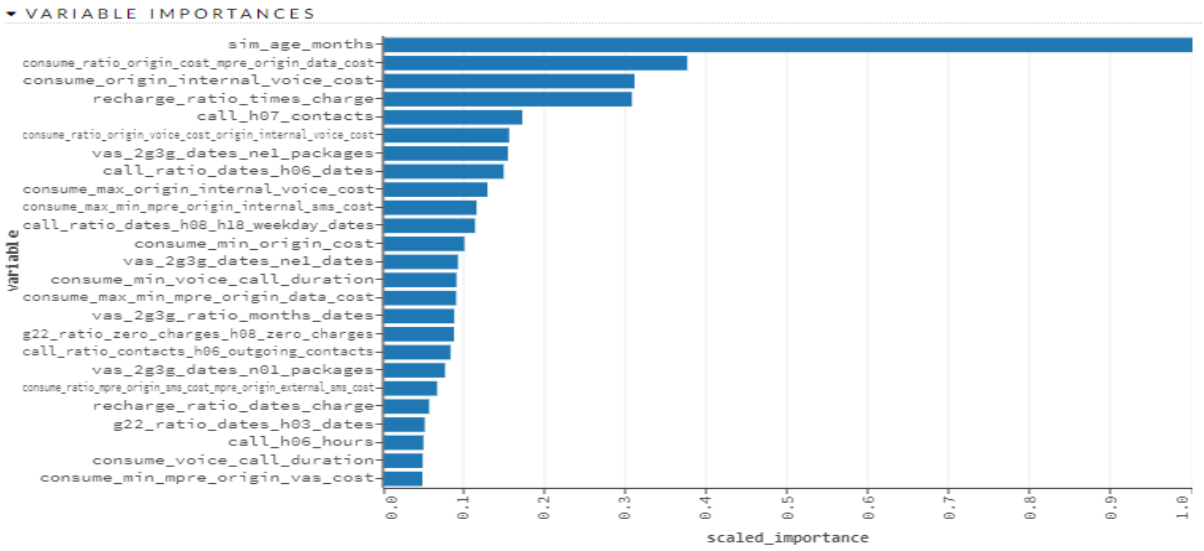
0	580	82	0.1239	82 / 662	0.88
1	82	551	0.1295	82 / 633	0.87
Total	662	633	0.1266	164 / 1,295	
Recall	0.88	0.87			

Chúng ta có thể thấy, mô hình XGBoost đã đoán trúng được 1131 mẫu trong tổng số 1295 quan sát, độ chính xác mà mô hình đạt được là 87,3%. Sử dụng các giá trị của bảng ma trận nhầm lẫn, chúng ta dựng nên biểu đồ đường cong ROC và tính ra được AUC = 93.8%.



Hình 3.3 Đường cong ROC cho mẫu kiểm thử

Như vậy, thuật toán XGBoost đã dựng nên một mô hình phân loại tốt. Bước tiếp theo chúng tôi sẽ xem xét các đặc trưng quan trọng nhất mà mô hình đã sử dụng.



Hình 3.4 Top các đặc trưng theo độ quan trọng

Có thể thấy, tuổi sim (sim_age_months) đang có trọng số cao nhất trong mô hình, điều này có lẽ đúng khi đa phần sinh viên đều là người mới sử dụng điện thoại, hoặc có thói quen thay đổi sim liên tục, không giống như lớp người đi làm, thường sẽ sử dụng cố định một số di động. Tiếp theo là tỉ lệ sử cước dụng dịch vụ giá trị gia tăng (vas) và cước dùng 3g (consume_ratio_origin_cost_mpre_origin_data_cost), tỉ lệ cước gọi nội mạng (consume_origin_internal_voice_cost) trên tổng cước thoại, tỉ lệ số lần nạp tiền trên tổng số tiền nạp (recharge_ratio_times_charge), số người gọi trong khung giờ 7h (call_h07_contacts) .. Các chỉ số này khá phù hợp với lịch trình và điều kiện sinh hoạt chung của sinh viên.

Sau khi xác định được các đặc trưng, chúng ta phải quay lại bước tổng hợp đặc trưng, xem các đặc trưng có phân bố hợp lý hay không, có đặc trưng nào bị thiên lệch. Sau khi loại bỏ các đặc trưng thiên lệch, chúng ta lại quay lại huấn luyện mô hình. Làm đi làm lại các bước nhiều lần, chúng ta sẽ đạt được mô hình tốt nhất để áp dụng dự đoán cho toàn bộ tập thuê bao.

Như vậy, ta có thể thấy phân bố các giá trị của đặc trưng là tương tự nhau giữa các nhãn. Điều này thể hiện các đặc trưng trên có thể đại diện cho sự phân hóa

của nhãn, hay có thể nói thuật toán đã hoạt động khá tốt, các đặc trưng lựa chọn đã khá chính xác, và chúng ta có thể lựa chọn thuật toán XGBoost để thực hiện công việc dự đoán trên toàn bộ tập dữ liệu.

3.4 Kết luận

Như vậy, sau quá trình thực nghiệm, tôi đã chọn ra được mô hình XGBoost là mô hình có chất lượng tốt nhất. Sử dụng mô hình này để áp dụng cho toàn bộ dữ liệu của nhà mạng, quá trình này cũng mất rất nhiều thời gian và công sức, vì khối lượng dữ liệu là vô cùng lớn, vì vậy hệ thống chạy rất lâu mới có kết quả. Sau khi đã phân loại được thuê bao, tôi đã thực hiện gọi điện để kiểm tra bằng tay. Kết quả cũng khá khả quan, khi tỉ lệ đạt là 80%. Vì vậy, trong thời gian tới, để nâng cao chất lượng sản phẩm, tôi sẽ tiếp tục tối ưu phần xử lý dữ liệu và xây dựng đặc trưng. Ngoài ra, dựa vào nền tảng có sẵn từ dự án này, tôi có thể phát triển sang các ngành nghề khác, các bài toán khác trong dự án bài toán dữ liệu lớn của nhà mạng.

KẾT LUẬN

Các nhà mạng viễn thông có một khối lượng dữ liệu lớn và đa dạng về cách hành vi sử dụng di động của khách hàng. Bằng các công cụ học máy hiện đại, chúng ta có thể trích xuất ra rất nhiều thông tin hữu ích từ đó, như chân dung khách hàng, thói quen, sở thích hoặc xu hướng của họ.

Đóng góp của báo cáo này là chúng tôi đã cho thấy sử dụng thuật toán XGBoost dựa trên một bộ các đặc trưng có tính phân lập mạnh mẽ từ các bản ghi thô của dữ liệu viễn thông phức tạp để phân loại các thuê bao có là sinh viên hay không. Chúng tôi đã đánh giá 811 đặc trưng này, và thấy rằng chúng có khả năng bao phủ cho khá nhiều mẫu đánh giá khác nhau. Từ đó, không chỉ là dự đoán rằng một thuê bao có là sinh viên hay không, chúng tôi có thể áp dụng cho việc dự đoán các nghề nghiệp khác tùy thuộc vào mẫu thu thập, hoặc chúng tôi còn có thể áp dụng các bài toán khác như đánh giá sở thích, điểm tín dụng cá nhân, điểm tích cực của một thuê bao…

Việc xác định các đặc trưng là tùy thuộc vào từng cá nhân, từng nhiệm vụ cụ thể, vì vậy ngoài các đặc trưng đã có, chúng tôi vẫn phải tiếp tục nghiên cứu dữ liệu và tìm hiểu thêm các đặc trưng mới. Bởi dữ liệu là vô cùng, và cách kết hợp dữ liệu là vô tận, nên định hướng tiếp theo của nghiên cứu chúng tôi vẫn là xây dựng đặc trưng mới, đánh giá hiệu quả mô hình trên thực tiễn và hiệu chỉnh mô hình khi cần thiết.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] https://moet.gov.vn/thong-ke/Pages/thong-ke.aspx - Truy cập ngày 20/05/2021

[2] https://vietnamcredit.com.vn/products/vietnam-industries/bao-cao-nganh-vien-thong-viet-nam-2020-54 – Truy cập ngày 20/05/2021

Tiếng Anh

[3] CE Shannon (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal* 27(3), 379–423.

[4] Chawla N (2005), "Data mining for imbalanced datasets: an overview", *Data mining and knowledge discovery handbook*, Berlin: Springer, Berlin, 853–867

[5] Yoav Ben-Shlomo, Sara Brookes, Matthew Hickman (2013). *Lecture Notes: Epidemiology, Evidence-based Medicine and Public Health*, 6th Edition, Wiley - Blackwell, Oxford.

[6] Fawcett, Tom (2006). "An Introduction to ROC Analysis", *Pattern Recognition Letters* 27 (8), 861–874

[7] Kuhn, Max; Johnson, Kjell (2013), *Applied Predictive Modeling*, NY: Springer, New York

[8] Ho, Tin Kam (1995), "Random Decision Forests", *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282.

[9] Jerome H. Friedman (2001), "Greedy function approximation: A gradient boosting machine.." *Ann. Statist* 29(5), 1189 - 1232.

[10] Powers, David M W (2011), "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", *Journal of Machine Learning Technologies* 2 (1), 37–63.

[11] Quinlan, J. R. (1986). "Induction of decision trees", *Machine Learning* 1(1), 81-106

[12] Tianqi Chen, Carlos Guestrin (2016), "XGBoost: A Scalable Tree Boosting System", "*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*", ACM, 785–794.