HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Quang Tuấn

MỘT SỐ THUẬT TOÁN HỌC MÁY TRONG PHÂN LOẠI HÀNH VI SỬ DỤNG GÓI CƯỚC DATA VIỄN THÔNG

CHUYÊN NGHÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 8.48.01.01

TÓM TẮT LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS TS. Trần Đình Quế**

Phản biện 1: PGS.TS. BÙI THU LÂM

Phản biện 2: PGS.TS. PHẠM VĂN CƯỜNG

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: 9 giờ ..... ngày 09 tháng 01 năm 2021

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

Ngày nay, trong kỷ nguyên kỹ thuật số, với sự bùng nổ của thông tin, số lượng dữ liệu do con người tạo ra ngày càng khổng lồ. Số lượng điện thoại smartphone và thiết bị kết nối tăng nhanh chóng, ngành công nghiệp viễn thông tràn ngập với số lượng dữ liệu khổng lồ. Nguồn gốc của số lượng dữ liệu khổng lồ này bao gồm dữ liệu lưu lượng truy cập mạng, mô hình sử dụng dữ liệu của khách hàng, dữ liệu vị trí, ứng dụng đã tải về,… Ngành công nghiệp viễn thông đang ngày càng thay đổi và phát triển không ngừng. Điện thoại thông minh đã trở thành một nhu cầu cơ bản của mỗi người trong cuộc sống ngày nay. Mọi người có thể kết nối với nhau ở bất cứ nơi nào trên thế giới, xóa bỏ rào cản khoảng cách. Mọi thông tin đều có thể được thu thập và xử lý nhanh hơn bao giờ hết. Và phân tích Big Data sẽ tạo điều kiện cho các ngành công nghiệp viễn thông phát triển mạnh mẽ trong thế giới kỹ thuật số. Các ứng dụng của phân tích số liệu trong lĩnh vực viễn thông, dữ liệu lớn là một cơ hội chuyển đổi ngành viễn thông sang hướng hoạt động hiệu quả hơn nhờ gia tăng mức độ hài lòng của khách hàng, tăng doanh thu nhờ tăng sản lượng và loại hình dịch vụ cung cấp, cắt giảm chi phí vận hành, giảm thiểu thiệt hại.

Trong khuôn khổ luận văn tập trung vào các kỹ thuật xử lý dữ liệu lớn và các thuật toán phân lớp dữ liệu bao gồm: Phân loại tuyến tính, Hồi quy logistic, Phân loại Naïve Bayes, Rừng ngẫu nhiên (RF). Ứng dụng thuật toán học máy trong lĩnh vực kinh doanh viễn thông sử dụng dữ liệu lịch sử của tập khách hàng để xây dựng các mô hình có khả năng phân loại, dự đoán nhu cầu sử dụng của khách hàng. Tập kết quả đó sẽ được dùng để hỗ trợ các đơn vị kinh doanh truyền thống đưa ra quyết định trong các chiến dịch kinh doanh của doanh nghiệp.

Cấu trúc của bài luận văn gồm 3 chương:

Chương 1: Tổng quan về bài toán phân loại hành vi sử dụng dịch vụ viễn thông: Trong chương này trình bày tổng quan quy trình phân tích dữ liệu, hệ thống xử lý dữ liệu phân tán và các phương pháp xử lý dữ liệu.

Chương 2: Mô hình hành vi và một số thuật toán học máy: Chương này sẽ đi sâu vào tìm hiểu 3 thuật toán là rừng ngẫu nhiên, phân loại Naïve Bayes, hồi quy Logistic.

Chương 3: Thử nghiệm và đánh giá: Chương này sẽ nêu mục tiêu thử nghiệm bài toán, ý nghĩa các chỉ số đo và thử nghiệm xây dựng mô hình dự đoán lần lượt với 3 thuật toán nêu trên và đánh giá kết quả.

CHƯƠNG 1: TỔNG QUAN VỀ BÀI TOÁN PHÂN LOẠI HÀNH VI SỬ DỤNG DỊCH VỤ VIỄN THÔNG

0.1. Giới thiệu bài toán

Các ứng dụng của phân tích số liệu trong lĩnh vực viễn thông, dữ liệu lớn là một cơ hội chuyển đổi ngành viễn thông sang hướng hoạt động hiệu quả hơn nhờ gia tăng mức độ hài lòng của khách hàng, tăng doanh thu nhờ tăng sản lượng và loại hình dịch vụ cung cấp, cắt giảm chi phí vận hành, giảm thiểu thiệt hại. Trong khuôn khổ luận văn tập trung vào các kỹ thuật xử lý dữ liệu lớn và các thuật toán phân lớp dữ liệu bao gồm: Phân loại tuyến tính, Hồi quy logistic, Phân loại Naïve Bayes, Rừng ngẫu nhiên (RF). Ứng dụng thuật toán học máy trong lĩnh vực kinh doanh viễn thông sử dụng dữ liệu lịch sử của tập khách hàng để xây dựng các mô hình có khả năng phân loại, dự đoán nhu cầu sử dụng của khách hàng. Tập kết quả đó sẽ được dùng để hỗ trợ các đơn vị kinh doanh truyền thống đưa ra quyết định trong các chiến dịch kinh doanh của doanh nghiệp.

0.2. Tổng quan quy trình phân tích dữ liệu

1.2.1. Tổng quan

- Sự kiện bắt đầu: Kinh doanh gửi PYC thực hiện dự án.

- Sự kiện kết thúc: Triển khai theo dõi kết quả và hành vi sau tác động.

- Đầu vào: Tài liệu đánh giá phạm vi mục tiêu của chương trình ứng dụng kinh doanh dựa trên phân tích dữ liệu.

- Đầu ra:

- Bảng dữ liệu sau quá trình mô hình dự đoán
- Chương trình kinh doanh tác động đến khách hàng cuối dựa trên phân tích dữ liệu.
- Báo cáo kết quả đánh giá chương trình.
- Triển khai mở rộng và xây dựng các chiến dịch định kỳ

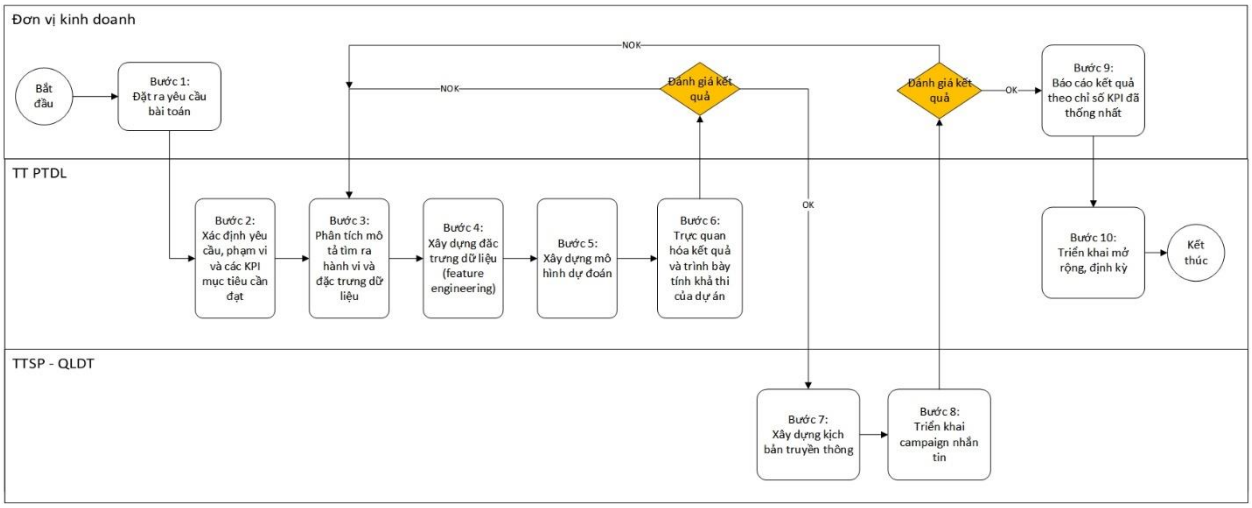
1.2.2. Quy trình triển khai bài toán phân tích dữ liệu

Bảng 1.1: Quy trình triển khai bài toán phân tích dữ liệu

Hoạt động chính	Các nội dung quan trọng
1. Đưa ra yêu cầu bài toán	Đơn vị kinh doanh: đưa ra yêu cầu bài toán, mô tả rõ hiện trạng và mục tiêu đầu ra mong muốn về cả doanh thu và tỷ lệ take up rate.
2. Xác định yêu cầu, phạm vi và các KPI mục tiêu cần đạt	Đơn vị kinh doanh: Đặt ra mục tiêu đầu ra mong muốn về cả doanh thu và tỷ lệ take up rate. Xác định các KPI và con số để đánh giá mô hình dự đoán trong bài toán phân tích. Xác định các KPI về kết quả triển khai của campaign ứng dụng phân tích dữ liệu.
3.Phân tích mô tả (Descriptive analytics)	Hypothesis testing Clean data, Explore Data, data understanding and preparation Plan phân tích Phân tích và chứng minh giả thiết
4.Xây dựng đặc trưng dữ liệu (feature engineering)	TT PTDL đưa ra danh sách đặc trưng liên quan đến dữ liệu. TTSP sử dụng kết quả trực quan hóa và kinh nghiệm về mặt kinh doanh giúp đóng vai trò tư vấn
5.Xây dựng model dự đoán phù hợp với chương trình (Predictive analystics)	TT PTDL xây dựng mô hình dự đoán theo các đặc trưng dữ liệu đã thống nhất.

Hoạt động chính	Các nội dung quan trọng
6.Trực quan hóa kết quả, thuyết phục với đơn vị ra yêu cầu	TT PTDL trực quan hóa đặc tính của các thuê bao được dự đoán. Thuyết phục đơn vị kinh doanh về kết quả đầu ra
7.Xây dựng kịch bản truyền thông	Lựa chọn sản phẩm, offer phù hợp với đặc tính từng nhóm thuê bao Xây dựng kịch bản tác động, nội dung tin nhắn, thời điểm, trigger tác động...
8.Triển khai và theo dõi kết quả	Phối hợp với các đơn vị P.QLDT, TTSP, TKCS để khai báo campaign tác động đến khách hàng cuối. Chia tập tác động thành 2 tập Target Group – để tác động và Control Group - để và theo dõi Xây dựng Dashboard để theo dõi các chỉ số KPI và diễn biến hành vi thuê bao sau tác động
9.Báo cáo kết quả	Báo cáo kết quả chương trình tới BTGĐ
10.Triển khai mở rộng, định kỳ	Nếu kết quả chương trình tốt, triển khai mở rộng và đựng thành luồng định kỳ hàng ngày/hàng tháng

1.2.3. Lưu đồ quy trình thực hiện dự án ứng dụng phân tích dữ liệu



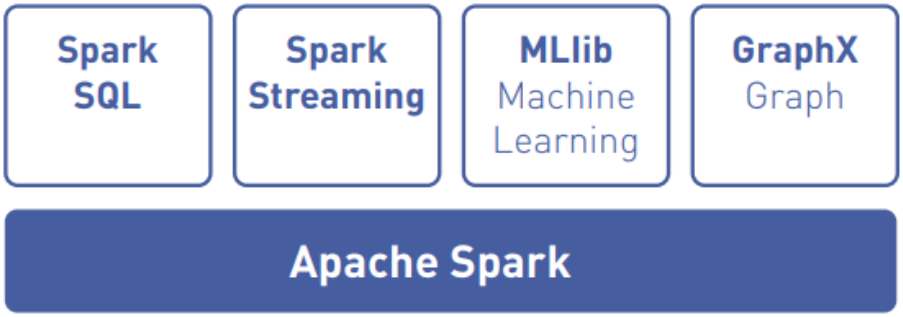
Hình 1.1 Lưu đồ quy trình thực hiện dự án ứng dụng phân tích dữ liệu

0.3. Xử lý dữ liệu phân tán với Spark

1.3.1. Giới thiệu

Thành phần chính của Spark là Spark Core: cung cấp những chức năng cơ bản nhất của Spark như lập lịch cho các tác vụ, quản lý bộ nhớ, fault recovery, tương tác với các hệ thống lưu trữ…Đặc biệt, Spark Core cung cấp API để định nghĩa RDD (Resilient Distributed DataSet) là tập hợp của các item được phân tán trên các node của cluster và có thể được xử lý song song.

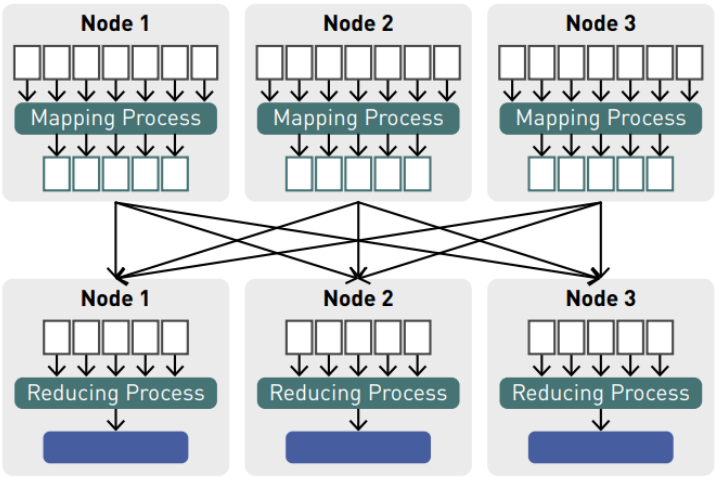
Spark có thể chạy trên nhiều loại Cluster Managers như Hadoop YARN, Apache Mesos hoặc trên chính cluster manager được cung cấp bởi Spark được gọi là Standalone Scheduler.



Hình 1.2: Các thành phần chính của Spark

1.3.2. Cơ chế hoạt động

Để tìm hiểu spark chúng ta sẽ bắt đầu với lịch sử hình thành và phát triển của nó. Trước Spark chúng ta đã từng biết tới MapReduce- một framework xử lý dữ liệu phân tán giúp Google thiết lập các index trong sự bùng nổ của nội dung web, trên các cụm máy chủ lớn.



Hình 1.3: Cơ chế hoạt động của ứng dụng Spark

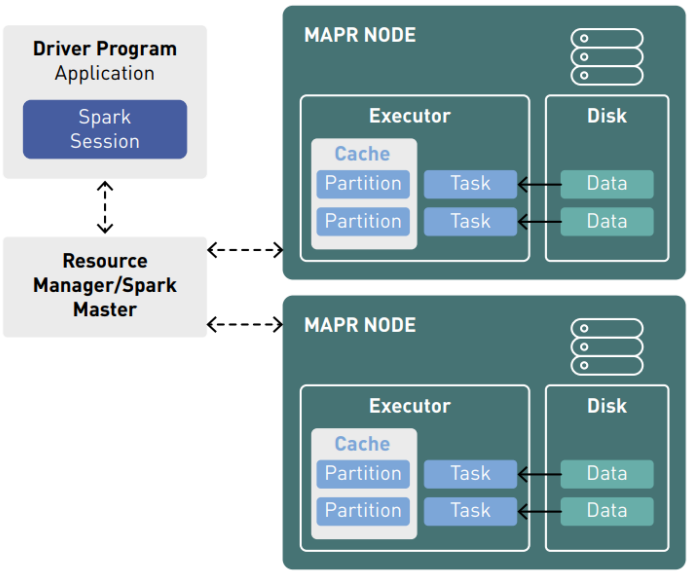
Có ba khái niệm cốt lõi trong chiến lược của Google:

- *Distribute Data*: Khi một tệp dữ liệu được tải lên cụm, nó sẽ được chia thành các phần được gọi là data block sau đó được phân phối chạy trên các data nodes và nhân rộng trên các cluster.
- *Distribute computation:* người dùng chỉ định map function để xử lý dữ liệu dựa trên các cặp key/value. Để tạo ra một tập các cặp key/value và kết hợp chúng với reduce function thì tất cả các giá trị trung gian được liên kết với cùng một khóa. Một chương trình được viết theo cấu trúc này sẽ tự động chạy song song trên 1 cụm cluster lớn.

1.3.3. Spark application

Biểu đồ bên dưới biểu diễn luồng chạy của một ứng dụng Spark chạy trên một cụm cluster.

- Mỗi ứng dụng spark chạy dưới dạng các quy trình độc lập được điều phối bởi Spark Session.
- Trình quản lý tài nguyên hay quản lý cluster sẽ phân công nhiệm vụ cho các worker, một task cho một partition.
- Mỗi task được giao cho 1 phần khối lượng của dataset trong partition của nó và output sẽ sẽ được xuất ra ở partition dataset mới.
- Kết quả được gửi trở lại driver application hoặc có thể được lưu vào ổ đĩa.



Hình 1.4: Luồng hoạt động của ứng dụng Spark

0.4. Các chỉ số đánh giá hiệu năng mô hình

1.4.1. Ma trận nhầm lẫn (Confusion matrix)

1.4.2. Các chỉ số Accuracy, Precision, Recall và F1 score

Accuracy: Chỉ số đánh giá độ chính xác tổng thể của mô hình. Giá trị của độ chính xác nằm trong khoảng 0 đến 1. Với 1 là giá trị độ chính xác tốt nhất và 0 là giá trị độ chính xác thấp nhất của một mô hình dự đoán. Độ chính xác (ACC) được tính bằng số tất cả các dự đoán đúng chia cho tổng số dự đoán của tập dữ liệu.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Chỉ số đánh giá tổng số dự đoán chính xác nhãn 1 chia cho tổng số dự đoán được dự đoán là nhãn 1. Giá trị lớn nhất của độ chính xác là 1 và nhỏ nhất là 0. Để tính Precision ta sử dụng công thức sau:

$$Prec = \frac{TP}{TP + FP}$$

Recall: Chỉ số thể hiện mô hình dự đoán đúng bao nhiêu phần nhãn 1 trong tổng số lượng nhãn 1 của cả tập. Nó còn có tên gọi là Tỉ lệ dương tính thực (TPR). Để tính recall ta sử dụng công thức sau:

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

F1-score: Chỉ số kết hợp giữa 2 chỉ số Precision và Recall. Để tính F1-score ta sử dụng công thức sau:

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Đường cong ROC

Đường cong ROC (receiver operating characteristic) là biểu đồ thể hiện hiệu năng phân loại nhãn của mô hình trên tất cả các ngưỡng điểm phân loại. Biểu đồ được tạo nên từ hai trục chứa giá trị True Positive Rate và False Positive Rate.

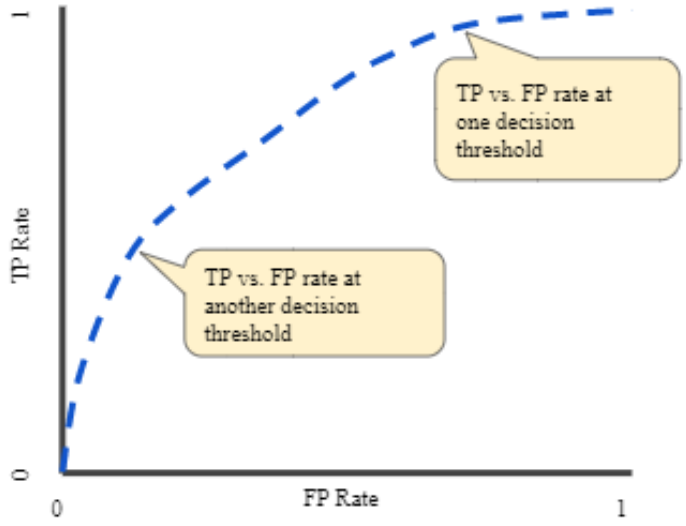
True Positive Rate (TPR) hay chính là Recall đã được trình bày ở phần trên. Công thức tính TPR:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) được tính bởi công thức:

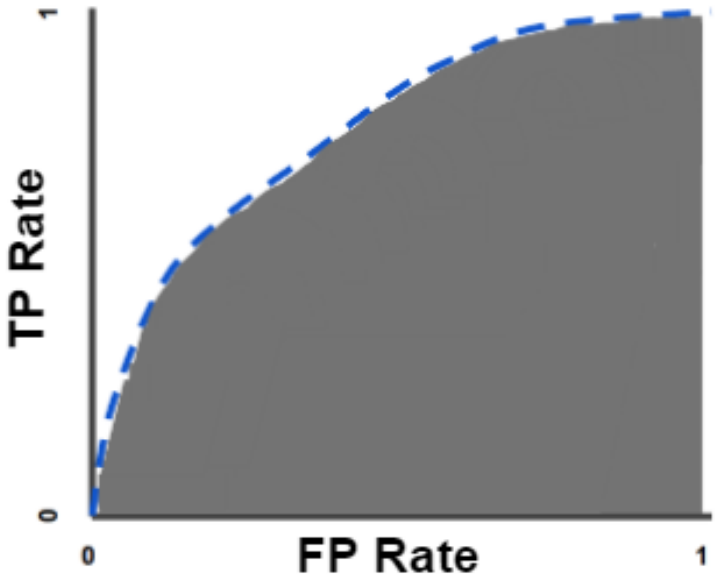
$$FPR = \frac{FP}{FP + TN}$$

Biểu đồ đường cong ROC được vẽ bởi các giá trị khác nhau của TPR và FPR trên mỗi ngưỡng cắt khác nhau của phân lớp. Việc hạ thấp ngưỡng phân loại sẽ phân loại nhiều được nhiều nhãn dương tính song cũng làm tăng cả đúng nhãn dương tính và sai nhãn dương tính.



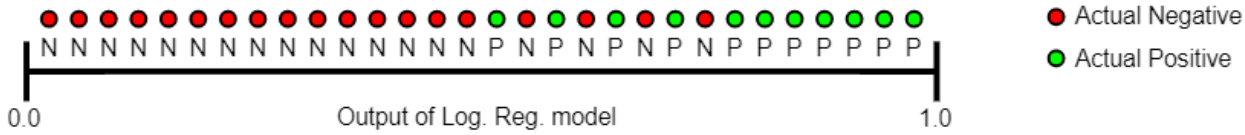
Hình 1.5: Đường cong ROC

Để đánh giá một mô hình người ta sử dụng AUC: Area Under the ROC Curve. AUC được tính bằng diện tích phần hình nằm bên dưới đường cong. Giá trị diện tích đó nằm trong khoảng [0,1].



Hình 1*Error! No text of specified style in document.****..6: Diện tích bên dưới đường cong ROC***

AUC là độ đo để đánh giá hiệu suất dự đoán trên tất cả các ngưỡng phân loại có thể có của mô hình dự đoán. Hay nói một cách khác thì AUC là xác suất mà mô hình xếp hạng một mẫu dương tính ngẫu nhiên cao hơn một mẫu âm tính ngẫu nhiên.



Hình 1*Error! No text of specified style in document.**.7: Xác suất phần loại nhãn***

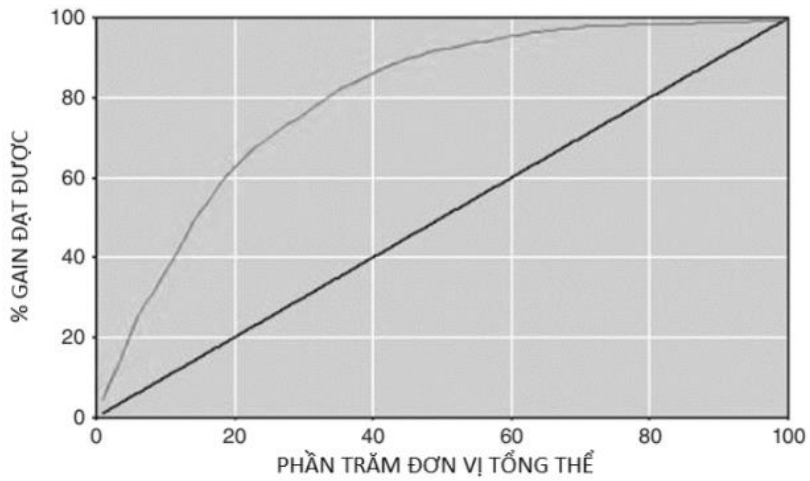
Hình trên mô phỏng một tập bản ghi được sắp xếp theo thứ tự tăng dần về điểm số xác suất phân loại nhãn. AUC có giá trị từ 0 đến 1. Một mô hình dự đoán sai 100% sẽ có AUC = 0 và dự đoán đúng 100% sẽ có AUC = 1. AUC có thể cho chúng ta thấy hiệu suất dự đoán của mô hình trên toàn bộ ngưỡng điểm do mô hình trả ra nhưng lại không cho ta biết ngưỡng điểm mô hình dự đoán chính xác nhất.

1.4.4. Biểu đồ Lift

Biểu đồ Gain kết hợp với biểu đồ Lift để thể hiện rõ hơn độ hiệu quả của mô hình phân loại. Biểu đồ Gain cung cấp cho chúng ta thông tin là trong % số đơn vị tổng thể chúng ta có thể đạt được bao nhiêu % đơn vị dữ liệu được phân loại chính xác.

1.4.5 Biểu đồ Gain

Biểu đồ Gain kết hợp với biểu đồ Lift để thể hiện rõ hơn độ hiệu quả của mô hình phân loại. Biểu đồ Gain cung cấp cho chúng ta thông tin là trong % số đơn vị tổng thể chúng ta có thể đạt được bao nhiêu % đơn vị dữ liệu được phân loại chính xác. Dựa trên kết quả tính toán ở bước xây dựng biểu đồ Lift chúng ta sẽ xây dựng được biểu đồ Gain tương ứng.



Hình 1 Error! No text of specified style in document.**.8: Biểu đồ Gain**

1.5 Các phương pháp xây dựng đặc trưng dữ liệu

Xây dựng đặc trưng dữ liệu là tiến trình lựa chọn các đặc tính của tập dữ liệu hay giảm số lượng các trường dữ liệu trong quá trình xây dựng các mô hình dự đoán. Với mục đích giảm thời gian tính toán, chi phí và cải thiện hiệu năng dự đoán của mô hình. Có nhiều phương pháp để lựa chọn đặc trưng dữ liệu nhưng có thể chia chúng thành ba nhóm chính:

- Phương pháp lọc: Xác định một số chỉ số nhất định và dựa trên các chỉ số đó để lựa chọn đặc trưng. Ví dụ như dựa vào chỉ số tương quan hoặc chi bình phương.
- Phương pháp đóng gói: Phương pháp này xem xét việc lựa chọn một tập các đặc trưng như một vấn đề tìm kiếm. Ví dụ như thuật toán đệ quy loại bỏ tính năng.
- Phương pháp nhúng: Phương pháp nhúng sử dụng các thuật toán có các phương pháp lựa chọn đặc trưng được tích hợp sẵn. Ví dụ như Lasso và RF có các phương pháp lựa chọn đặc trưng riêng của nó.

1.5.1 Các phương pháp thống kê lựa chọn đặc trưng dữ liệu với phương pháp lọc

Các phương pháp thống kê lựa chọn đặc trưng dữ liệu với phương pháp lọc thường sử dụng các chỉ số thể hiện mức độ tương quan giữa các biến đầu vào và biến đầu ra để làm cơ sở cho việc lựa chọn đặc trưng. Do đó việc lựa chọn các phương pháp thống kê phụ thuộc nhiều vào kiểu dữ liệu của các biến. Các kiểu dữ liệu phổ biến bao gồm dữ liệu dạng số và dữ liệu dạng phân loại, mỗi loại có thể chia thành nhiều kiểu dữ liệu như dạng số nguyên, dạng số thập phân cho dữ liệu dạng số và dạng nhị phân, thứ tự và định danh cho dữ liệu dạng phân loại.

1.5.1.1 Hệ số tương quan Pearson’s

Hệ số tương quan là một chỉ số thống kê đo mối liên hệ tương quan giữa hai biến số. Giá trị của hệ số tương quan r ($-1 \leq r \leq 1$). Hệ số tương quan càng gần 0 hoặc bằng 0 có nghĩa là hai biến đang xét không có mối liên hệ gì với nhau; ngược lại nếu giá trị của hệ số tương quan càng gần 1 hoặc -1 nghĩa là hai biến có mối quan hệ tuyệt đối. Nếu hệ số tương quan có giá trị âm thì đó là hai biến nghịch biến và hệ số tương quan dương thì đó là hai biến đồng biến. Hiện nay có nhiều công thức để tính hệ số tương quan giữa hai biến nhưng thông dụng nhất là công thức tính hệ số tương quan Pearson. Tương quan Person sẽ xác định một đường thẳng phù hợp nhất với mối quan hệ tuyến tính của hai biến. Xét hai biến số x và y được lấy từ n mẫu, hệ số tương quan Pearson sẽ được tính bằng công thức sau:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

1.5.1.2 Hệ số tương quan hạng Spearman

Hệ số tương quan hạng Spearman được sử dụng thay thế hệ số tương quan Pearson để kiểm tra mối quan hệ giữa hai biến được xếp hạng hoặc một biến được xếp hạng và một biến đo lường. Sử dụng khi phân phối của tổng thể được giả sử không phải là phân phối chuẩn hoặc trong trường hợp có các giá trị quan sát bất thường (lớn quá hoặc nhỏ quá).

$$spearman_{cor} = 1 - \frac{6\sum_i^n d_i^2}{n(n^2-1)}$$

Trong đó d_i là hiệu hạng của 2 biến được tính bằng:

$$d_i = rankX_i - rankY_i$$

1.5.1.3 Kiểm định chi bình phương (Chi squared)

Là phương pháp tính hệ số tương quan giữa các biến độc lập và biến phụ thuộc. Các biến được chọn làm đặc trưng của tập dữ liệu là các biến có hệ số Chi bình phương lớn. Công thức tính Chi bình phương:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Trong đó: O_i là các giá trị quan sát

E_i là các giá trị kỳ vọng

1.5.2 Các phương pháp thống kê lựa chọn đặc trưng dữ liệu với phương pháp đóng gói

Đệ quy loại bỏ đặc tính (Recursive Feature Elimination-RFE) là một trong những phương pháp lựa chọn đặc trưng dữ liệu phổ biến nhất hiện nay. RFE sẽ loại bỏ các trường dữ liệu có tương quan yếu đối với biến phụ thuộc cho tới khi đạt tới số lượng trường dữ liệu cần thiết do người dùng xác định từ trước. Với số lượng trường dữ liệu ít hơn mô hình dự đoán sẽ chạy hiệu quả hơn, giảm tài nguyên, thời gian chạy và đôi khi là nâng cao hiệu năng dự đoán. RFE hoạt động bằng cách tìm kiếm một tập con các trường dữ liệu bắt đầu bằng việc sử dụng tất cả các trường dữ liệu. Sau mỗi lần huấn luyện mô hình, các trường dữ liệu sẽ được

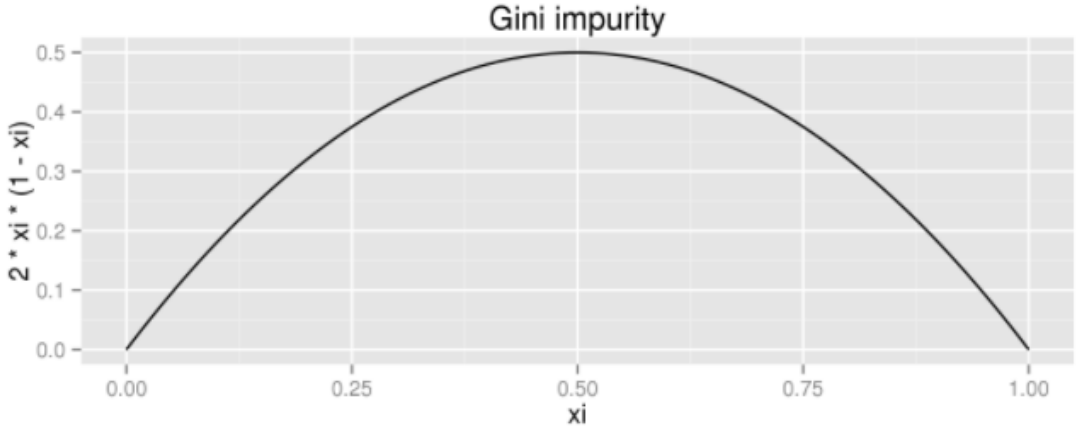
sắp xếp theo thứ tự giảm dần của mức độ quan trọng. Sau đó các trường dữ liệu mức độ quan trọng thấp sẽ được bỏ ra và lặp lại quá trình huấn luyện.

1.5.3 Các phương pháp thống kê lựa chọn đặc trưng dữ liệu với phương pháp nhúng

Sử dụng thuật toán Rừng ngẫu nhiên để tính mức độ quan trọng của các thuộc tính. Đối với thuật toán rừng ngẫu nhiên mỗi lần thực hiện phân chia tại nốt cha sẽ tạo ra hai lớp con có chỉ số độ thuần khiết GINI nhỏ hơn nốt cha.

Công thức tính độ thuần khiết GINI:

$$G = \sum_{i=1}^{n} p_i(1 - p_i)$$



Hình 1.9: Đồ thị biểu diễn độ thuần khiết GINI

Tại mỗi nốt chỉ số đánh giá mức độ quan trọng của thuộc tính sẽ được tính bằng công thức:

$$I = G_{parent} - G_{split1} - G_{split2}$$

Trong đó: G_{parent} là độ thuần khiết của nốt cha

G_{split1} là độ thuần khiết của nốt con thứ nhất

G_{split2} là độ thuần khiết của nốt con thứ hai

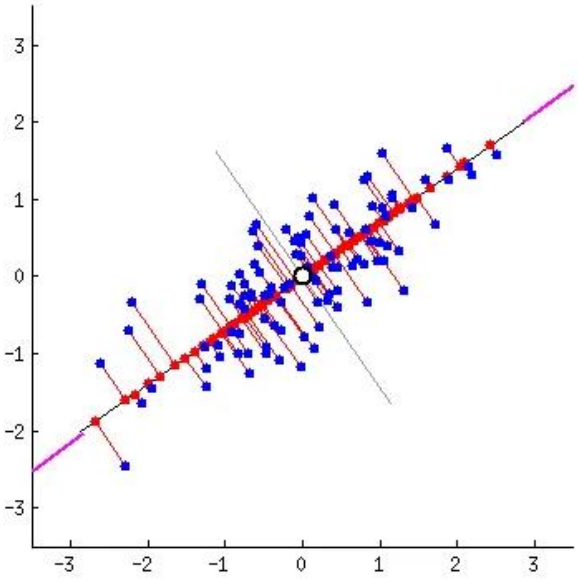
1.6 Kỹ thuật tiền xử lý dữ liệu

Kỹ thuật tiền xử lý dữ liệu là một trong những kỹ thuật tối quan trọng trong quá trình xây dựng các mô hình dự đoán với các thuật toán học máy. Chúng ta đều biết rằng các thuật toán học máy sẽ dựa vào tập dữ liệu đầu vào để đưa ra kết quả dự đoán. Nhưng vấn đề lớn

nhất mà các mô hình này gặp phải là chất lượng dữ liệu đầu vào không đủ tốt. Đó chính là lý do chúng ta dành phần lớn thời gian trong quá trình xây dựng mô hình dự đoán cho tiến trình tiền xử lý dữ liệu. Các kỹ thuật tiền xử lý dữ liệu là điểm khác biệt lớn giữa mô hình dự đoán tốt và mô hình dự đoán không tốt.

1.7 Thuật toán giảm chiều dữ liệu (PCA)

Thuật toán giảm chiều dữ liệu PCA (Principal Components Analysis) là kỹ thuật chuyển đổi các trường dữ liệu trong tập dữ liệu thành các trường dữ liệu mới gọi là Principal Component (PCs). Mục tiêu chính là số trường dữ liệu mới giảm tối thiểu nhất có thể so với số lượng trường dữ liệu ban đầu mà vẫn chứa đủ những thông tin đại diện cho cả tập dữ liệu. Hay nói cách khác PCA là kỹ thuật gộp các trường dữ liệu hiện hành. Mỗi trường dữ liệu mới là tổ hợp có trọng số của các trường dữ liệu gốc. Các PC được hình thành theo cách gán trọng số lớn hơn cho các PC thành phần có tính đại diện lớn hơn cho dữ liệu gốc.



Hình 1.10: Mô phỏng thuật toán PCA

Kết chương

CHƯƠNG 2: MÔ HÌNH HÀNH VI VÀ MỘT SỐ THUẬT TOÁN HỌC MÁY

2.1 Thuật toán rừng ngẫu nhiên (Random Forest)

2.1.1 Cây quyết định

2.1.1.2 Khái niệm

Cây quyết định (Decision tree) là một mô hình supervised learning, có thể được áp dụng vào cả hai bài toán classification và regression. Việc xây dựng một decision tree trên dữ liệu huấn luyện cho trước là việc đi xác định các câu hỏi và thứ tự của chúng. Decision tree có thể làm việc được với tập dữ liệu có đặc trưng dạng categorical và dạng numerical. Decision Tree là thuật toán có cấu trúc dạng cây, trong đó mỗi internal node thể hiện cho một thuộc tính dữ liệu, mỗi nhánh con của node biểu diễn giá trị của thuộc tính và mỗi leaf node sẽ chứa class label.

2.1.1.2 Ý tưởng thuật toán

Bước 1: Bắt đầu với việc set tập dữ liệu S ở root node

Bước 2: Lặp lại việc tính toán Entropy(H) và Information Gain(IG) với từng thuộc tính

Bước 3: Lựa chọn thuộc tính có Entropy nhỏ nhất hoặc Information Gain lớn nhất làm internal node

Bước 4: Chia tập S theo từng thuộc tính đã được lựa chọn để tạo ra các tập con dữ liệu

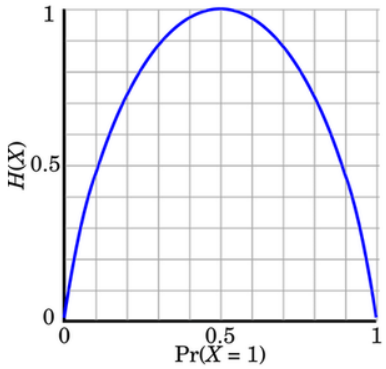
Bước 5: Thuật toán lặp lại trên mỗi tập con và chỉ xem xét các thuộc tính chưa được lựa chọn làm internal node trước đó.

2.1.1.3 Cơ sở lý thuyết

a. Hàm số Entropy

Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x1,x2,\ldots,xn$. Giả sử rằng xác suất để x nhận các giá trị này là $pi=p(x=xi)$. Ký hiệu phân phối này là $\mathbf{p}=(p1,p2,\ldots,pn)$. Entropy của phân phối này là:

$$\mathrm{H}_{(p)} = -\sum_{i=1}^{n} p_i \, log_2^{p_i}$$



Hình 1.11: Đồ thị của hàm Entropy

b. Information Gain

Information Gain được tính dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất. Do H(S) là không đổi với mỗi tầng, ta chọn thuộc tính f có Entropy nhỏ nhất để thu được Gain(x,S) lớn nhất.

$$G_{(x,S)} = H_{(s)} - H_{(x,S)}$$

Trong đó: H(S) là Entropy tổng của toàn bộ tập data set S.

H(x,S) là Entropy được tính trên thuộc tính x.

2.1.2 Thuật toán rừng ngẫu nhiên (Random Forest)

2.1.2.1 Khái niệm

Random forest là một tập hợp các mô hình (ensemble) gồm nhiều cây quyết định (decision tree). Mô hình Random Forest rất hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thể mạnh yếu khác nhau, nhưng theo nguyên tắc "wisdom of the crowd", ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

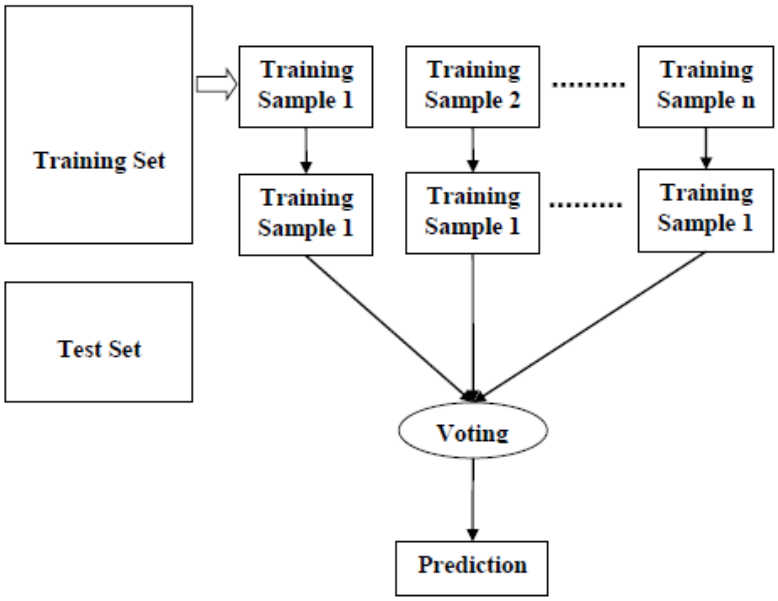
Như tên gọi của nó, Random Forest (RF) dựa trên cơ sở :

- Random = Tính ngẫu nhiên
- Forest = nhiều cây quyết định (decision tree)

Đơn vị của RF là thuật toán cây quyết định, với số lượng hàng trăm. Mỗi cây quyết định được tạo ra một cách ngẫu nhiên từ việc: Tái chọn mẫu (bootstrap, random sampling) và chỉ dùng một phần nhỏ tập biến ngẫu nhiên (random features) từ toàn bộ các biến trong dữ

liệu. Ở trạng thái sau cùng, mô hình RF thường hoạt động rất chính xác, nhưng đổi lại, rất khó để có thể hiểu được cơ chế hoạt động bên trong mô hình vì cấu trúc quá phức tạp.

2.1.2.2 Ý tưởng thuật toán



Hình 1.12: Ý tưởng thuật toán Rừng ngẫu nhiên

2.1.2.3 Ưu điểm, nhược điểm

2.2 Thuật toán Naïve Bayes

2.2.1 Suy diễn Bayes

Suy diễn Bayes là một phương pháp suy diễn thống kê, trong đó định lý Bayes được sử dụng để cập nhật xác suất/khả năng xảy ra của một giả thuyết khi càng nhiều dữ liệu/thông tin về giả thuyết đó được cung cấp đầy đủ. Suy diễn Bayes được hình thành dựa trên xác suất có điều kiện. Biết rằng A và B là hai sự kiện xảy ra, khi đó xác xuất xảy ra A với điều kiện B biết trước được tính bằng công thức:

$$P(A|B) = \frac{P(A|B)\,P(A)}{P(B)}$$

Trong đó: P(A|B): là xác suất hậu nghiệm (posterior probability)

P(B|A): là xác suất hợp lý (likelihood probability)

P(A): là xác suất tiên nghiệm (prior probability)

P(B): là thực chứng (evidence)

2.2.2 Cơ sở lý thuyết

Đối với các bài toán phân loại trong machine learning, phương pháp Naïve-Bayes được dùng tương đối phổ biến và đem lại kết quả khả quan. Trong thuật toán này, xác suất có điều kiện được ứng dụng để xác định xác suất xảy ra tại từng nhãn và chọn ra nhãn có xác suất cao nhất với điều kiện là các trường dữ liệu features của một điểm dữ liệu. Giả sử thuật toán phân loại Naïve-bayes chỉ ra nhãn Y cho bởi các điểm dữ liệu, x1, x2, …xn và xác suất hậu nghiệm trong suy diễn Bayes (coi theta Θ là Y, và data là x1,x2,..xn) với xác suất xảy ra như sau:

$$P(Y|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|Y)P(Y)}{P(x_1, x_2, \ldots, x_n)}$$

$$P(\boldsymbol{\Theta}|data) = \frac{P(data|\boldsymbol{\Theta}) \times P(\boldsymbol{\Theta})}{P(data)}$$

2.2.3 Ứng dụng của Bayes trong phân tích dữ liệu

Trong các bài toàn phân tích dữ liệu, trường phân loại (nhãn) trong tập dữ liệu gốc thường không bao gồm đầy đủ cho các điểm dữ liệu. Do vậy, việc phân tích và ước lượng xác suất trên một tập sample (tập có đầy đủ nhãn) và suy đoán trên toàn tập lớn (population) là hoàn toàn cần thiết.

2.3 Thuật toán Logistic Regression

2.3.1 Khái niệm

Logistic Regression (Hồi quy logistic) là một mô hình hồi quy nhằm dự đoán phân lớp giá trị đầu ra ứng với một vector đầu vào. Nói cách khác, mục tiêu phương pháp nhằm phân loại các đối tượng vào các lớp tương ứng. Đầu vào của mô hình là một tập dữ liệu với các biến phụ thuộc và biến độc lập. Mô hình sẽ sử dụng giá trị của các biến phụ thuộc để dự đoán giá trị của biến độc lập. Đối với bài toán Logistic regression thì đầu ra của bài toán là xác suất dự đoán ứng với từng giá trị của biến độc lập.

2.3.2 Cơ sở lý thuyết

Sử dụng phương pháp thống kê ta có thể cho rằng khả năng một đối tượng có các thuộc tính x nằm vào một nhóm y0 là xác suất của nhóm y0 khi biết x: $p(y_0|x)$

Dựa vào công thức xác suất có điều kiện ta có:

$$p(y_0|x) = \frac{p(x\,|y_0)\,p(y_0)}{p(x)} = \frac{p(x\,|y_0)\,p(y_0)}{p(x\,|y_0)\,p(y_0) + p(x\,|y_1)\,p(y_1)}$$

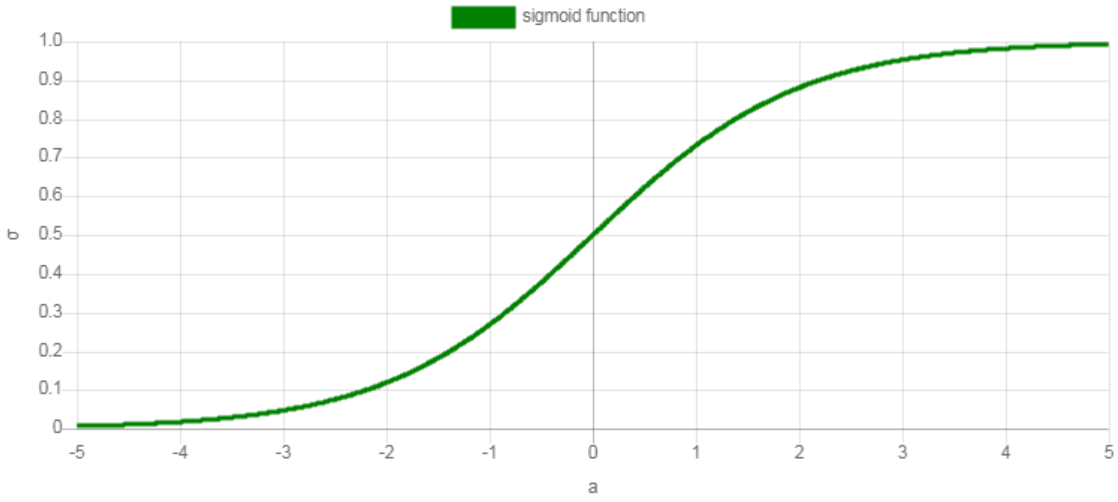
Nếu ta đặt:

$$a = \ln \frac{p(x\,|y_0)\,p(y_0)}{p(x\,|y_1)\,p(y_1)}$$

Ta có:

$$p(y_0|x) = \frac{1}{1+e^{-a}} = \sigma(a)$$

Hàm ở trên được gọi là hàm sigmoid của biến a, khi vẽ phân phối của a và hàm sigmoid, ta có:



Hình 1. Đồ thị hàm sigmoid σ(a)

Hình 1*Error! No text of specified style in document.****.13: Đồ thị hàm sigmoid***

Kết chương

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1 Đặt vấn đề

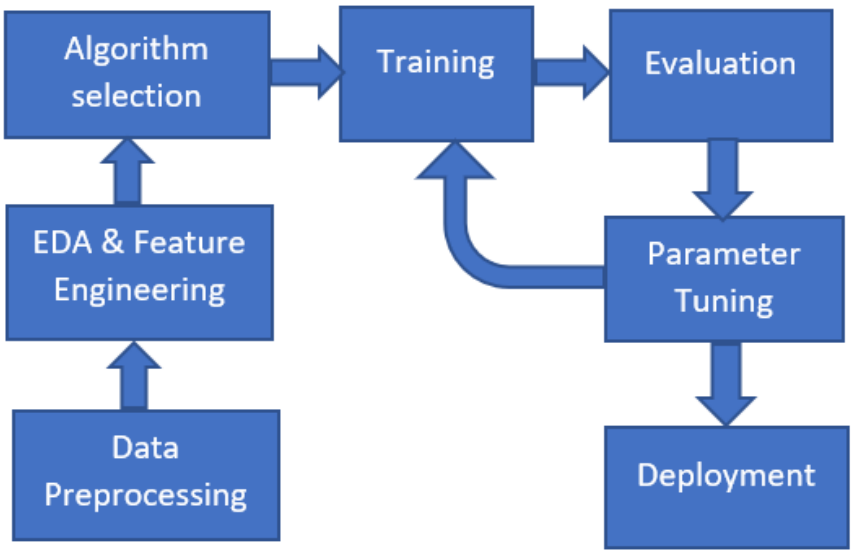
Ứng dụng thuật toán học máy trong lĩnh vực kinh doanh viễn thông sử dụng dữ liệu lịch sử của tập khách hàng để xây dựng các mô hình có khả năng phân loại, dự đoán nhu cầu sử dụng của khách hàng. Tập kết quả đó sẽ được dùng để hỗ trợ các đơn vị kinh doanh truyền thống đưa ra quyết định trong các chiến dịch kinh doanh của doanh nghiệp.

3.2 Xác định bài toán

Mục tiêu bài toán: Xây dựng mô hình dự đoán tập khách hàng có nhu cầu sử dụng gia tăng về lưu lượng, tiêu dùng dịch vụ. Song song với đó là xây dựng mô hình đề xuất sản phẩm viễn thông phù hợp với nhu cầu gia tăng tiêu dùng của khách hàng. Thử nghiệm xây dựng mô hình dự đoán lần lượt với 3 thuật toán là Hồi quy tuyến tính, Phân loại Naïve Bayes và Rừng ngẫu nhiên (RF). Từ đó so sánh hiệu năng để tìm ra thuật toán phù hợp nhất với bộ dữ liệu đang xét. Sau đó ứng dụng kết quả dự đoán của mô hình vào thực tế so sánh hiệu quả dựa trên các chỉ số và tỉ lệ dự đoán đúng tự nhiên.

3.3 Quy trình xây dựng mô hình học máy

Quy trình xây dựng một mô hình học máy cơ bản sẽ gồm các bước sau:



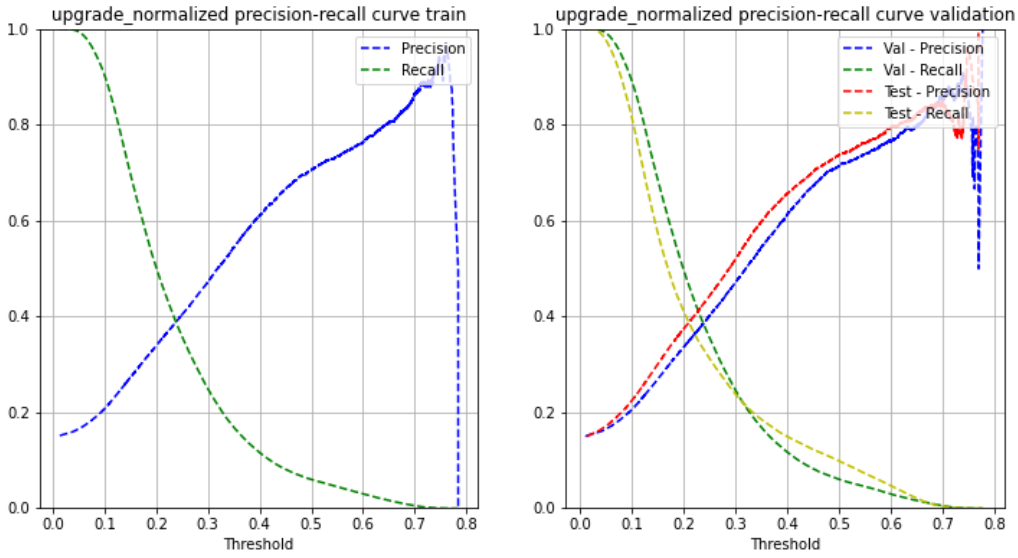
Hình 1 Error! No text of specified style in document.**.14: Các bước xây dựng mô hình học máy**

3.4 Thực nghiệm

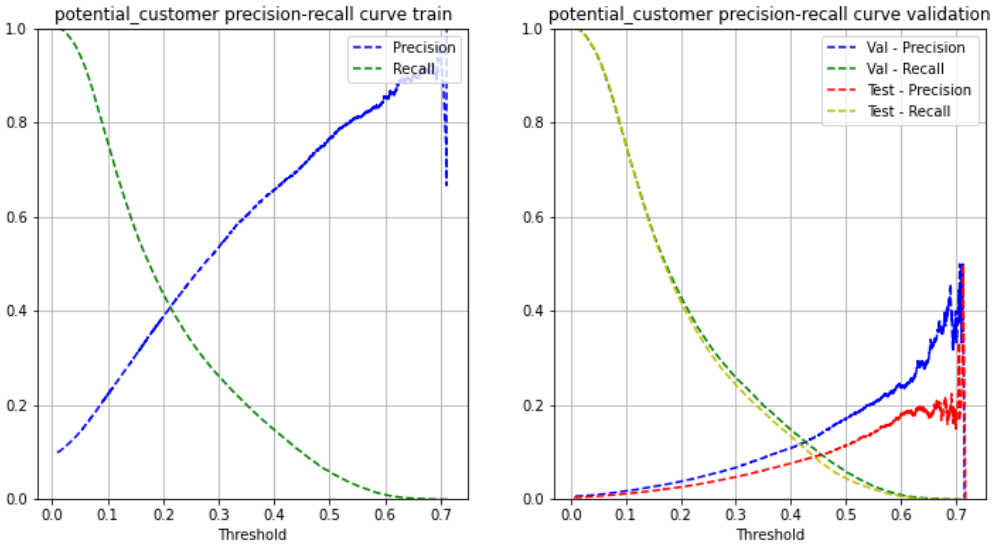
3.5 Kết quả thực nghiệm

Trong khuôn khổ bài luận văn em đã thử nghiệm xây dựng 3 mô hình dự đoán nhu cầu dùng tăng dịch vụ data của nhà mạng Viettel. Với cùng một bộ dữ liệu huấn luyện mô hình bao gồm các dữ liệu liên quan tới lịch sử sử dụng data, gọi thoại, nhắn tin, nạp tiền của các thuê bao sử dụng dịch vụ viễn thông của nhà mạng Viettel. Cả ba tập dữ liệu đầu vào 3 mô hình này là giống nhau và cùng được tiền xử lý dữ liệu như nhau để đảm bảo công bằng trong việc so sánh hiệu năng dự đoán của các mô hình.

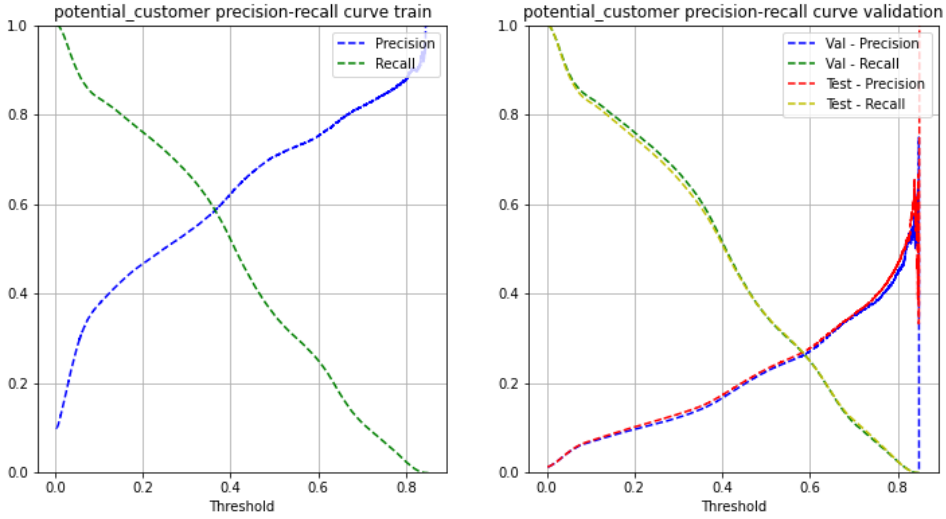
Biểu đồ biểu diễn độ đo precision và recall thể hiện tỉ lệ dự đoán đúng và độ phủ của mô hình trên hai tập dữ liệu huấn luyện và xác thực. Từ mỗi điểm trên hình ta sẽ xác định được ứng với từng phần trăm của tập dữ liệu dự đoán thì sẽ có độ chính xác và độ phủ là bao nhiêu. Nếu phần trăm của tập dữ liệu càng lớn thì độ chính xác càng giảm và độ phủ càng tăng.



Hình 1.15: Precision-Recall thuật toán Naïve Bayes



Hình 1.16: Precision-Recall thuật toán hồi quy Logistic



Hình 1.17: Precision-Recall thuật toán rừng ngẫu nhiên

3.6 Xây dựng hệ thống

3.6.1 Giới thiệu hệ thống

Mục đích: Xây dựng hệ thống quản lý machine learning work flow, bao gồm toàn bộ các quá trình training, testing, inference 1 data scientist thường phải thực hiện.

Một số mục tiêu cụ thể:

- Cung cấp giao diện quản lý các tiến trình training, testing, inference
- Cung cấp giao diện quản lý các model machine learning
- Cơ chế sử dụng AutoML để tự động chọn ra model và bộ tham số tối ưu
- Cơ chế manual define tham số và training nhiều model đồng thời

Các chức năng chính:

- Tạo mới model:

- Tạo mới các model với 2 chế độ: người dùng tự define tham số, hoặc sử dụng AutoML tự động chọn lựa tham số
- Khi tạo mới model xong có thể chọn đồng thời các option để train/ test hoặc inference
- Chọn dataset tương ứng với các option trên
- Chọn location trên HDFS để lưu trữ report và output

- Sử dụng model:

- Chọn model trong list model đã được train
- Khi chọn model xong có thể chọn đồng thời các option để train/ test hoặc inference
- Chọn dataset tương ứng với các option trên
- Chọn location trên HDFS để lưu trữ report và output

- Quản lý các model:

- Hệ thống lưu lại và quản lý các model đã được train
- Hệ thống hiển thị chi tiết của từng model (tham số, training score, validation score, ...)
- Hệ thống lưu lại logs, report cho các lần chạy

3.6.2 Biểu đồ ca sử dụng hành vi người dùng

3.6.3 Biểu đồ ca sử dụng giám sát dự án

3.6.4 Biểu đồ ca sử dụng giám sát mô hình

3.6.5 Giao diện Home

3.6.6 Giao diện thanh điều hướng

3.6.7 Giao diện thông tin chung

3.6.8 Giao diện nguồn dữ liệu

3.6.9 Giao diện thông tin mô hình

3.7 Kết quả trong triển khai thực tế

3.7.1 Các chỉ số tính hiệu quả triển khai

3.7.2 Kết quả triển khai thực tế

Kết chương

KẾT LUẬN CHUNG

Các kết quả thu được trong luận văn

Sau khi thử nghiệm lần lượt 3 thuật toán phân loại Naïve Bayes, hồi quy Logistic, rừng ngẫu nhiên đối trên cùng một tập dữ liệu thì thuật toán rừng ngẫu nhiên cho kết quả dự đoán chính xác nhất sau đó tới thuật toán hồi quy Logistic và thuật toán Naïve Bayes cho kết quả dự đoán kém chính xác nhất.

Hiện nay các bài toán mô hình upgrade tương đối tốt và đã đạt tới ngưỡng gần như không thể improve bằng các kĩ thuật thông thường mà chỉ có thể improve dựa trên việc xây dựng các features có giá trị phân loại tốt hơn; thêm vào đó cách xây dựng mô hình upgrade tương đối đơn giản nên trước mắt chưa có phương pháp để improve mô hình này. Thêm vào đó, việc mất cân đối giữa số lượng các sản phẩm trong mô hình cũng ảnh hưởng lớn đến performance chung của mô hình khi đưa ra dự đoán cho các sản phẩm thiểu số.

Định hướng nghiên cứu tiếp theo

Dựa trên những vấn đề đó định hướng hiện tại đang thử nghiệm các phương pháp sau:

- Từ bộ p1 thu được tương ứng với từng sản phẩm, giả định đầu tiên là với score p1 cao hơn tương ứng với thuê bao có khả năng mua sản phẩm ấy cao hơn. Chọn ra sản phẩm có p1 cao nhất để tiến hành back test. Kết hợp kết quả back test với phân tích cluster.
- Từ kết quả thu được của trial 1st cũng như từ performance của các model riêng biệt. Đánh giá rằng các mô hình thu được hoạt động tốt trong khả năng dự đoán xu hướng sử dụng gói của thuê bao, tuy nhiên để đưa đến một xếp hạng score cuối cùng rằng sản phẩm nào sẽ được mua thì cần một phương pháp rõ ràng và chính xác hơn. Vì vậy, thử áp dụng các mô hình phân lớp cho bộ score p1 thu được từ các model trên với mục tiêu là khi đó có thể xây dựng được một "MODEL MASTER" có khả năng tổng hợp kết quả từ các mô hình nhỏ.

Từ kết quả quan sát tại các thử nghiệm trước cũng như quá trình xây dựng mô hình riêng biệt, nhận thấy: Tuy các model đều bị ảnh hưởng bởi hiện tượng imblance và đã sử dụng các phương pháp Downsampling để cải thiện performance. Nhưng các yếu tố có thể ảnh hưởng đến ranking scores của các mô hình không chỉ là tỉ lệ nhãn giữa 0 và 1 (0:1) mà còn một yếu tố chưa được đề cập tới đó chính là số lượng tuyệt đối của các nhãn trong mô hình. Bởi số lượng các thuê bao có mua gói ngày/n-ngày so với số lượng các thuê bao mua gói tháng có số lượng chênh lệch đáng kể.

Vì vậy, để calibrate thành công kết quả các mô hình cần một hàm có khả năng cover được tất cả các yếu tố trên. Từ đó, tiến hành nghiên cứu tìm hiểu các phương pháp calibrate khác thường được áp dụng. Kết quả là một số phương pháp phổ biến như Platt’s Scaling và Isotonic Regression đã được đề cập đến trong những tài liệu calibration khác.