ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN

ĐỀ TÀI

Ứng dụng Data Mining để dự đoán khách hàng hủy thuê bao trong ngành viễn thông

GV hướng dẫn: TS. Nguyễn Huy Trường

Sinh viên thực hiện: Lê Quang Đức - 20227221

Hà Nội, Ngày 17 tháng 5 năm 2025

Mục lục

Mở đầu

Lý do chọn đề tài

Trong thị trường viễn thông cạnh tranh, giữ chân khách hàng quan trọng hơn bao giờ hết, vì chi phí tìm khách hàng mới thường cao gấp nhiều lần so với duy trì khách hàng hiện tại. Việc dự đoán sớm khách hàng có nguy cơ hủy thuê bao giúp doanh nghiệp triển khai các biện pháp giữ chân hiệu quả.

Nhờ sự phát triển của Data Mining và Machine Learning, doanh nghiệp có thể phân tích dữ liệu để xác định nhóm khách hàng rời bỏ tiềm năng, từ đó tối ưu chiến lược chăm sóc khách hàng. Vì vậy, em chọn đề tài này nhằm ứng dụng các phương pháp phân tích dữ liệu để hỗ trợ doanh nghiệp giảm tỷ lệ hủy thuê bao và nâng cao hiệu quả kinh doanh.

Đối tượng và phạm vi nghiên cứu

1. Đối tượng nghiên cứu trong báo cáo này là bộ dữ liệu dự đoán khách hàng có khả năng hủy đăng ký thuê bao dựa trên những hành vi của khách hàng. Bộ dữ liệu được lấy từ trên trang Kaggle chứa thông tin cung cấp dịch vụ Internet của một công ty viễn thông ở Canada.

2. Phạm vi nghiên cứu là các khái niệm liên quan đến bài toán phân loại nhị phân; các khái niệm, đánh giá bài toán và các bước xây dựng mô hình phân loại hồi quy logistic

Tóm tắt nội dung

- **Chương 1: Đặt vấn đề:** Tổng quan về ngành dịch vụ viễn thông và bài toán dự đoán khách hàng rời bỏ ngành dịch vụ.
- **Chương 2: Cơ sở lý thuyết**: Trình bày khái niệm về bài toán phân loại. Tìm hiểu mô hình hồi quy logistic.
- **Chương 3. Ứng dụng các phương pháp phân loại vào bài toán dự đoán khách hàng rời bỏ**: Tìm hiểu và tiến hành các bước khai phá dữ liệu. Xây dựng và đánh giá mô hình

Chương 1

Đặt vấn đề

1.1 Tổng quan về ngành dịch vụ viễn thông

Ngành viễn thông đóng vai trò quan trọng trong nền kinh tế số, cung cấp hạ tầng kết nối cho cá nhân, doanh nghiệp và chính phủ. Viễn thông bao gồm các dịch vụ như điện thoại cố định, di động, internet băng thông rộng, truyền hình cáp và các giải pháp công nghệ số.

Ngành dịch vụ viễn thông thường có các đặc điểm sau:

1. **Tính cạnh tranh cao:** Các nhà mạng liên tục đưa ra các gói cước và chương trình ưu đãi để thu hút và giữ chân khách hàng.
2. **Chi phí đầu tư lớn:** Hạ tầng viễn thông đòi hỏi đầu tư mạnh vào công nghệ và cơ sở hạ tầng mạng.
3. **Mức độ đổi mới nhanh:** Công nghệ thay đổi liên tục, đòi hỏi doanh nghiệp phải thích ứng nhanh để không bị tụt hậu.
4. **Tỷ lệ khách hàng rời bỏ (Churn Rate) cao:** Do tính cạnh tranh, khách hàng dễ dàng chuyển đổi giữa các nhà cung cấp dịch vụ.

Ngành viễn thông đóng vai trò quan trọng trong nền kinh tế số, cung cấp hạ tầng kết nối cho cá nhân và doanh nghiệp. Với sự phát triển của 5G, AI, IoT, ngành này không chỉ cung cấp dịch vụ liên lạc mà còn mở rộng sang dữ liệu lớn, điện toán đám mây và thanh toán số. Cạnh tranh cao và đổi mới liên tục buộc doanh nghiệp phải tối ưu vận hành và nâng cao trải nghiệm khách hàng. Xu hướng chính gồm triển khai 5G, chuyển đổi số và mở rộng dịch vụ số hóa, giúp viễn thông trở thành trung tâm của nền kinh tế hiện đại.

1.2 Mục tiêu nghiên cứu

Trong bối cảnh ngành viễn thông ngày càng cạnh tranh, việc giữ chân khách hàng trở thành một thách thức lớn đối với doanh nghiệp. Tỷ lệ khách hàng hủy thuê bao (churn) không chỉ ảnh hưởng đến doanh thu mà còn làm tăng chi phí tìm kiếm khách hàng mới. Vì vậy, doanh nghiệp cần một hệ thống dự đoán chính xác khả năng rời bỏ của khách hàng để có biện pháp can thiệp kịp thời.

Bài toán đặt ra là làm thế nào để xây dựng một mô hình phân tích dữ liệu hiệu quả, dựa trên các thuật toán Machine Learning, giúp xác định các yếu tố ảnh hưởng đến churn và đưa

ra dự báo chính xác. Điều này sẽ hỗ trợ doanh nghiệp đề xuất các chiến lược giữ chân khách hàng phù hợp, tối ưu hóa dịch vụ và nâng cao trải nghiệm người dùng.

Trong đồ án này, em xin phép trình bày thuật toán hồi quy logistic giúp dự đoán phân loại khách hàng có khả năng rời bỏ ngành dịch vụ dựa theo những thông tin cá nhân của khách hàng.

Bộ dữ liệu sử dụng thu nhập được ở trên trang Kaggle cung cấp thông tin cá nhân của khách hàng đăng ký dịch vụ của một công ty viễn thông.

customerID	gender	SeniorCitizen	Partner	Dependen	tenure	PhoneSer	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentM	MonthlyCh	TotalChar	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed che	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed che	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-	Yes	Electronic	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit car	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed che	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed che	49.95	587.45	No
7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service	No internet service	No interne	No interne	No interne	No interne	Two year	No	Credit car	18.95	326.8	No
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-	Yes	Bank trans	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic	105.5	2686.05	No
3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service	No internet service	No interne	No interne	No interne	No interne	One year	No	Mailed che	20.65	1022.95	No
9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit car	55.2	528.35	Yes
4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic	90.05	1862.9	No

Chương 2

Cơ sở lý thuyết

2.1 Tổng quan về khai thác dữ liệu và học máy

Data Mining (khai phá dữ liệu) là quá trình khám phá và trích xuất thông tin, tri thức hữu ích từ một lượng lớn dữ liệu. Mục tiêu của Data Mining là tìm ra các mẫu (patterns), mối quan hệ (relationships), xu hướng (trends) hoặc luật (rules) tiềm ẩn trong dữ liệu để phục vụ cho việc ra quyết định, dự đoán hoặc tối ưu hóa các hoạt động trong thực tế. Data Mining là một phần quan trọng trong lĩnh vực Khoa học dữ liệu (Data Science) và Trí tuệ nhân tạo (AI).

Các kỹ thuật trong khai thác dữ liệu bao gồm:

- Phân loại (Classification): Dự đoán một nhãn cụ thể (ví dụ: khách hàng rời bỏ hay không).
- Phân cụm (Clustering): Nhóm các đối tượng có đặc điểm giống nhau.
- Khai thác luật kết hợp (Association Rules Mining): Tìm ra các mối quan hệ thường xảy ra cùng nhau trong dữ liệu.
- Dự báo (Prediction/Regression): Dự đoán giá trị trong tương lai.
- Phát hiện bất thường (Anomaly Detection): Phát hiện những điểm dữ liệu bất thường (outliers).

Machine Learning (ML)là một nhánh của trí tuệ nhân tạo, cho phép hệ thống "học" từ dữ liệu để cải thiện hiệu suất mà không cần lập trình rõ ràng từng bước.

Trong bài toán dự đoán khách hàng rời bỏ, mô hình học máy sẽ học từ dữ liệu lịch sử để phát hiện các đặc điểm chung của những khách hàng có khả năng rời bỏ, từ đó đưa ra dự đoán cho các khách hàng hiện tại.

2.2 Giới thiệu kỹ thuật phân loại

Bài toán phân loại (hay phân lớp) thuộc nhóm các phương pháp thống kê với một hoặc nhiều biến số vào quá trình tách các tập đối tượng hoặc quan sát đôi một phân biệt nhau và phân loại một đối tượng hoặc quan sát mới vào các lớp đã có sẵn. Cụ thể:

- **Mô tả dữ liệu (data description)**: đặc tả bằng hình ảnh (với số chiều không quá 3) hoặc bằng cấu trúc đại số các đặc trưng khác biệt giữa hai hay nhiều quần thể đã có sẵn. Ta cần tìm các "giá trị đặc trưng" (discriminant) sao cho trị số của các giá trị đó ở từng quần thể là khác nhau nhất có thể

- **Phân loại dữ liệu (data allocation)**: sắp xếp các đối tượng hoặc quan sát vào hai hay nhiều nhóm khác nhau. Nói cách khác, ta cần phát triển các "quy tắc" để phân loại các đối tượng một cách tối ưu, tức là có ít đối tượng bị phân loại sai nhất.

Có nhiều bài toán phân loại dữ liệu như phân loại nhị phân (*binary classification*), phân loại đa lớp (*multiclass classification*).

- **Bài toán phân loại nhị phân** là bài toán gán nhãn dữ liệu cho đối tượng vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (*feature*) của bộ phân lớp.
- **Bài toán phân loại đa lớp** là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn 2. Như vậy với từng dữ liệu phải xem xét và phân lớp chúng vào những lớp khác nhau chứ không phải là hai lớp như bài toán phân loại nhị phân. Và thực chất bài toán phân loại nhị phân là một bài toán đặc biệt của phân loại đa lớp.

Trong đồ án này, em sẽ tập trung đi vào nghiên cứu bài toán phân loại hai lớp, một bài toán phổ biến trong đời sống khi kết luận một đối tượng có hay không thỏa mãn một điều kiện nào đó.

2.3 Bài toán phân loại nhị phân

Phân loại nhị phân (*binary classification*) là bài toán phân loại các phần tử từ một tập hợp các đối tượng thành hai nhóm dựa trên cơ sở là chúng có một thuộc tính nào đó hay không (hay còn gọi là tiêu chí). Một số ví dụ điển hình cho bài toán phân loại nhị phân có thể kể đến như:

- Dự đoán xem khách hàng có khả năng rời bỏ dịch vụ không.
- Xác định xem một sản phẩm làm ra là đủ tốt để bán chưa, hay nên loại bỏ nó (thuộc tính để phân loại là tính đủ tốt).
- Xác định xem một người có hay không mua sản phẩm (thuộc tính để phân loại tình trạng của sản phẩm).

Thông thường, các bài toán phân loại nhị phân liên quan đến một lớp là trạng thái bình thường và một lớp khác là trạng thái bất thường. Ví dụ “không rời bỏ” là trạng thái bình thường và “rời bỏ” là trạng thái bất thường. Lớp cho trạng thái bình thường được gán nhãn lớp 0 và lớp có trạng thái bất thường được gán nhãn lớp 1.

Người ta thường lập mô hình nhiệm vụ phân lớp nhị phân với một mô hình dự đoán phân phối xác suất Bernoulli cho mỗi bài toán. Phân phối Bernoulli là một phân phối xác suất rời rạc bao gồm trong một sự kiện sẽ có kết quả nhị phân là 0 hoặc 1. Đối với phân lớp, điều này có nghĩa là mô hình dự đoán xác suất của một mẫu thuộc lớp 1, hoặc trạng thái bất thường.

Một số thuật toán phổ biến có thể sử dụng để phân loại nhị phân bao gồm:

- Hồi quy logistic.
- KNN (K-Nearest Neighbors).
- Cây quyết định (Decision Tree)

- SVM (Support Vector Machine)

Một số thuật toán được thiết kế đặc biệt để phân lớp nhị phân và không hỗ trợ nhiều hơn hai lớp, điển hình là hồi quy logistic. Tiếp theo, ta sẽ tìm hiểu cách đánh giá mô hình phân loại nhị phân.

2.4 Các chỉ số đánh giá mô hình

Đánh giá mô hình giúp chúng ta lựa chọn được mô hình phù hợp nhất đối với bài toán của mình. Tuy nhiên để tìm được thước đo đánh giá mô hình phù hợp thì chúng ta phải hiểu khái niệm, bản chất và trường hợp áp dụng của từng thước đo.

Giả định rằng chúng ta đang xây dựng một mô hình phân loại nợ xấu. Nhãn của các quan sát sẽ bao gồm GOOD (thông thường) và BAD (nợ xấu). Kích thước của các tập dữ liệu như sau:

- Tập train: 1000 hồ sơ bao gồm 900 hồ sơ GOOD và 100 hồ sơ BAD.
- Tập test: 100 hồ sơ bao gồm 85 hồ sơ GOOD và 15 hồ sơ BAD.

Để thuận tiện cho diễn giải và đồng nhất với những tài liệu tham khảo khác về ký hiệu thì biến mục tiêu y nhãn BAD có giá trị 1 và GOOD giá trị 0. Đồng thời trong các công thức đo lường thống kê, nhãn BAD là positive và GOOD là negative. **Positive** và **Negative** ở đây chỉ là quy ước tương ứng với giá trị 1 và 0 chứ không nên hiểu theo nghĩa đen là *tích cực* và *tiêu cực*.

Một mô hình phân loại đưa ra kết quả dự báo trên tập train được thống kê trên bảng chéo như sau:

Predict/Actual		Actual	
		BAD(Positive)	**GOOD(Negative**
Predict	**BAD (Positive)**	55 (TP - True Positive)	50 (FP - False Positive)
	GOOD (Negative)	45(FN - False Negative)	850(TN - True Negative)
Total		**100**	**900**

Bảng 2.1: Confusion Matrix

Các chỉ số TP, FP, TN, FN lần lượt có ý nghĩa là:

- TP (True Positive): Tổng số trường hợp dự báo khớp Positive.
- TN (True Negative): Tổng số trường hợp dự báo khớp Negative.
- FP (False Positive): Tổng số trường hợp dự báo các quan sát thuộc nhãn Negative thành Positive.
- FN (False Negative): Tổng số trường hợp dự báo các quan sát thuộc nhãn Positive thành Negative.

Những chỉ số trên sẽ là cơ sở để tính toán những metric như accuracy, precision, recall, f1 score.

2.4.1 Accuracy

Khi xây dựng mô hình phân loại chúng ta sẽ muốn biết một cách khái quát tỷ lệ các trường hợp được dự báo đúng trên tổng số các trường hợp là bao nhiêu. Tỷ lệ đó được gọi là accuracy (độ chính xác). Độ chính xác giúp ta đánh giá hiệu quả dự báo của mô hình trên một bộ dữ liệu. Độ chính xác càng cao thì mô hình của chúng ta càng chuẩn xác. Độ chính xác được tính qua công thức:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Tổng số mẫu}} \tag{2.1}$$

Trong ví dụ này sẽ là:

$$\text{Accuracy} = \frac{55 + 850}{1000} = 90.5\%$$

Khi đó ta nói mô hình f dự báo chính xác 90.5%

Trong các chỉ số đánh giá mô hình phân loại, độ chính xác (accuracy) thường được ưa chuộng nhờ công thức rõ ràng và ý nghĩa dễ hiểu. Tuy nhiên, nhược điểm của chỉ số này là đánh giá dựa trên toàn bộ các nhãn mà không phản ánh độ chính xác riêng biệt của từng nhãn. Vì vậy, nó không phù hợp với những bài toán mà tầm quan trọng của các nhãn không đồng đều. Chẳng hạn, trong bài toán phân loại nợ xấu, việc phát hiện chính xác một hồ sơ nợ xấu có ý nghĩa lớn hơn nhiều so với việc xác định đúng một hồ sơ bình thường.

Trong trường hợp này, ta cần tập trung đánh giá độ chính xác trên nhãn 1 (nợ xấu), và những chỉ số như precision và recall sẽ trở nên cần thiết vì chúng cung cấp cái nhìn chuyên sâu hơn vào hiệu quả dự đoán trên nhóm nhãn quan trọng này.

2.4.2 Precision

Precision trả lời cho câu hỏi trong các trường hợp được dự báo là positive thì có bao nhiêu trường hợp là đúng ? Và tất nhiên precision càng cao thì mô hình của chúng ta càng tốt. Công thức của precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.2}$$

Trong ví dụ này sẽ là:

$$\text{Precision} = \frac{55}{55 + 50} = 52.4\%$$

Precision cho chúng ta biết mức độ chính xác của mô hình trong việc dự báo các hồ sơ là nợ xấu. Cụ thể, khi precision = 52.4%, điều này có nghĩa là trong số các hồ sơ được mô hình dự đoán là nợ xấu, có 52.4% thực sự là nợ xấu. Một chỉ số khác cũng đánh giá hiệu suất dự báo trên nhóm dương tính (Positive) là recall. Mặc dù recall có cùng tử số với precision trong công thức tính (số lượng dự đoán đúng là nợ xấu), nhưng điểm khác biệt nằm ở mẫu số. Nếu như precision đo lường tỷ lệ đúng trên các dự đoán là nợ xấu, thì recall lại đo lường tỷ lệ đúng trên tổng số hồ sơ thực sự là nợ xấu. Nói cách khác, recall phản ánh khả năng phát hiện các trường hợp nợ xấu của mô hình.

2.4.3 Recall

Recall đo lường tỷ lệ dự báo chính xác các trường hợp Positive trên toàn bộ các mẫu thuộc nhóm Positive. Công thức của recall như sau:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3}$$

Trong ví dụ này sẽ là:

$$\text{Precision} = \frac{55}{55+45} = 55\%$$

Để tính được recall thì chúng ta phải biết trước nhãn của dữ liệu. Do đó recall có thể được dùng để đánh giá trên tập train vì chúng ta đã biết trước nhãn. Trên tập test khi dữ liệu được coi như mới hoàn toàn và chưa biết nhãn thì chúng ta sẽ sử dụng precision.

Mối quan hệ giữa precision và recall: Thông thường các model sẽ lựa chọn một ngưỡng mặc định là 0.5 để quyết định nhãn. Tức là nếu ta có một hàm phân loại f thì nhãn dự báo sẽ dựa trên độ lớn của xác suất dự báo như sau:

$$f(x) = \begin{cases} 1, & \text{nếu } f(x) \geq 0.5 \\ 0, & \text{nếu } f(x) < 0.5 \end{cases}$$

Do đó precision và recall sẽ không cố định mà chịu sự biến đổi theo ngưỡng xác suất được lựa chọn.

Ta còn có thể chứng minh được mối quan hệ giữa precision và recall khi biến đổi theo threshold là mối quan hệ đánh đổi (trade off). Khi precision cao thì recall thấp và ngược lại. Thật vậy:

- Giả sử trong bài toán phân loại nợ xấu, nếu chúng ta mong muốn mô hình chỉ đưa ra dự đoán khi thật sự "chắc chắn" một hồ sơ là nợ xấu, thì có thể thiết lập một ngưỡng xác suất cao, chẳng hạn 0.9. Khi đó, chỉ những hồ sơ mà mô hình dự đoán có xác suất từ 90% trở lên mới được phân loại là nợ xấu. Việc này giúp đảm bảo rằng các hồ sơ bị gắn nhãn nợ xấu gần như chắc chắn là nợ xấu thực sự — tức precision sẽ tăng. Tuy nhiên, do ngưỡng cao khiến mô hình trở nên "khắt khe", số lượng hồ sơ được dự đoán là nợ xấu sẽ giảm, dẫn đến việc bỏ sót nhiều trường hợp nợ xấu thực sự. Do đó, recall sẽ có xu hướng giảm.

- Ngược lại, nếu chúng ta muốn "nới lỏng" tiêu chí phân loại nợ xấu bằng cách giảm ngưỡng (ví dụ xuống 0.3 hoặc 0.4), mô hình sẽ đánh giá nhiều hồ sơ hơn là nợ xấu. Điều này làm tăng số lượng hồ sơ được dự đoán là nợ xấu, qua đó recall có thể tăng vì nhiều trường hợp nợ xấu thực sự được phát hiện hơn. Tuy nhiên, do việc "nới lỏng" này cũng kéo theo nhiều dự đoán sai (những hồ sơ bình thường bị dự đoán nhầm là nợ xấu), nên precision sẽ giảm vì tỷ lệ dự đoán đúng trong số các dự đoán nợ xấu bị loãng đi.

Sự đánh đổi giữa precision và recall khiến cho kết quả của mô hình thường sẽ là: precision cao, recall thấp hoặc precision thấp, recall cao. Khi đó rất khó để lựa chọn đâu là một mô hình tốt vì không biết rằng đánh giá trên precision hay recall sẽ phù hợp hơn. Chính vì vậy chúng ta sẽ tìm cách kết hợp cả precision và recall trong một chỉ số mới, đó chính là F1 score.

2.4.4 F1 score

F1 score là trung bình điều hoà giữa precision và recall. Do đó nó đại diện hơn trong việc đánh giá độ chính xác trên đồng thời precision và recall. Công thức của F1 score được tính bằng:

$$\text{F1} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}. \tag{2.4}$$

Trong trường hợp precision = 0 hoặc recall = 0, ta quy ước F1 = 0.

Ta chứng minh được rằng giá trị của F1 score luôn nằm trong khoảng của precision và recall. Thật vậy:

$$\begin{aligned}\text{F1} &= \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \\ &\leq \frac{2 \times (\text{precision} \times \text{recall})}{2\min(\text{precision}, \text{recall})} \\ &= \max(\text{precision}, \text{recall}),\end{aligned}$$

$$\begin{aligned}\text{F1} &= \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \\ &\geq \frac{2 \times (\text{precision} \times \text{recall})}{2\max(\text{precision}, \text{recall})} \\ &= \min(\text{precision}, \text{recall}).\end{aligned}$$

Do đó đối với những trường hợp mà precision và recall quá chênh lệch thì score sẽ cân bằng được cả hai độ lớn này và giúp ta đưa ra một đánh giá khách quan hơn.

2.4.5 ROC-AUC

ROC là đường cong biểu diễn khả năng phân loại của một mô hình phân loại tại các ngưỡng threshold. Đường cong này dựa trên hai chỉ số :

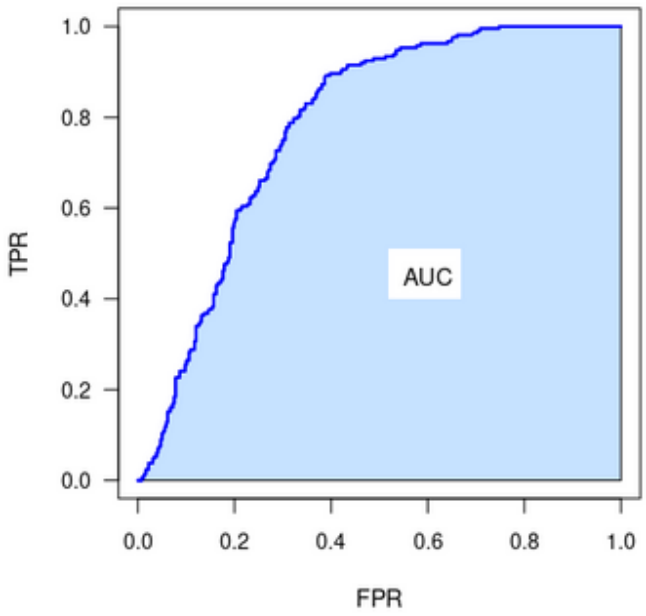
- **TPR** (true positive rate): Hay còn gọi là *recall* hoặc *sensitivity*. Là tỷ lệ các trường hợp phân loại đúng positive trên tổng số các trường hợp thực tế là positive. Chỉ số này sẽ đánh giá mức độ dự báo chính xác của mô hình trên nhóm positive. Khi giá trị của nó càng cao, mô hình dự báo càng tốt **trên nhóm positive**. Nếu TPR = 0.9, chúng ta tin rằng 90% các mẫu thuộc nhóm positive đã được mô hình phân loại đúng.

$$\text{TPR/recall/sensitivity} = \frac{TP}{\text{total positive}}$$

- **FPR** (false positive rate): Tỷ lệ dự báo sai các trường hợp thực tế là negative thành positive trên tổng số các trường hợp thực tế là negative. Nếu giá trị của FPR = 0.1, mô hình đã dự báo sai 10% trên tổng số các trường hợp là negative. Một mô hình có FPR càng thấp thì mô hình càng chuẩn xác vì sai số của nó **trên nhóm negative** càng thấp. Phần bù của FPR là specificity do lượng tỷ lệ dự báo đúng các trường hợp negative trên tổng số các trường hợp thực tế là negative.

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{\text{total negative}}$$

Đồ thị ROC là một đường cong cầu lồi dựa trên TPR và FPR có hình dạng như bên dưới:



Hình 2.1: Chỉ tiêu đánh giá ROC-AUC

AUC là chỉ số được tính toán dựa trên đường cong ROC (receiving operating curve) nhằm đánh giá khả năng phân loại của mô hình tốt như thế nào ? Phần diện tích gạch chéo nằm dưới đường cong ROC và trên trục hoành là AUC (area under curve) có giá trị nằm trong khoảng [0, 1]. Khi diện tích này càng lớn thì đường cong ROC có xu hướng tiệm cận đường thẳng $y = 1$ và khả năng phân loại của mô hình càng tốt. Khi đường cong ROC nằm sát với đường chéo đi qua hai điểm (0, 0) và (1, 1), mô hình sẽ tương đương với một phân loại ngẫu nhiên.

2.5 Mô hình hồi quy logistic

Trong phần này, tôi sẽ đề cập một phương pháp được sử dụng phổ biến trong bài toán phân loại nhị phân bởi đây là một mô hình cơ bản và có độ chính xác cao, đó chính là mô hình hồi quy logistic.

2.5.1 Giới thiệu

Nhắc lại hồi quy tuyến tính

Chúng ta đã quen thuộc mô hình tuyến tính có dạng:

$$y = f(w^T x) \tag{2.5}$$

Trong đó $f()$ là hàm kích hoạt (activation function) và x được hiểu là dữ liệu mở rộng với $x_0 = 1$ được thêm vào để thuận tiện cho việc tính toán. Trong hồi quy tuyến tính, tích vô hướng $w^T x$ được trực tiếp sử dụng để dự đoán đầu ra y , loại này phù hợp nếu chúng ta cần dự đoán một giá trị thực của đầu ra không bị chặn trên và dưới. Nhiệm vụ của chúng ta là tìm bộ tham số tối ưu (optimized parameters) phù hợp để tính toán.

Ngược lại, với bài toán hồi quy logistic (logistic regression), đầu ra dự đoán của nó sẽ nhận giá trị 0 hoặc 1, tham số tối ưu được huấn luyện dựa trên dữ liệu mang tính định tính. Qua đó, ta có thể thấy rằng hồi quy logistic là bài toán thuộc lớp các bài toán phân loại nhị phân (binary classification).

Ví dụ 2.5.1. *Xét một tập hai lớp, ta có thể đánh nhãn nữ là 1 và nam là 0, p là xác suất người đó là nữ, khi đó trung bình và phương sai là:*

$$\text{mean} = 0 \times (1-p) + 1 \times p = p$$
$$\text{variance} = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p).$$

Có thể thấy rõ ràng phương sai của biến không phải là hằng số. Với $p = .5$, nó sẽ bằng $.5 \times .5 = 25$ trong khi đó với $p = 8$, nó lại bằng $.8 \times .2 = .16$. Phương sai sẽ tiệm cận về 0 thay vì tiệm cận về 0 hoặc 1.

Đặt biến phản hồi $\hat{y}$ có kết quả là 0 hoặc 1. Nếu chúng ta lập mô hình xác suất bằng 1 bằng một mô hình tuyến tính dự đoán duy nhất, chúng ta có thể viết

$$p(x) = E(\hat{y} \mid x) = \beta_0 + \beta_1 x + \varepsilon.$$

Tuy nhiên, mô hình trên có một số nhược điểm:

- Giá trị dự đoán $\hat{y}$ có thể lớn hơn 1 hoặc nhỏ hơn 0 vì biểu thức tuyến tính cho giá trị kỳ vọng là không bị chặn.
- Phương sai của giá trị dự đoán $\hat{y}$ không đổi.

Qua ví dụ trên, ta thấy cần có cách tiếp cận mới mà phương sai của giá trị dự đoán là không được cố định và làm cho hàm mục tiêu $p(x)$ có điều kiện dựa trên các biến quan sát $\mathbf{X} = \mathbf{x}$.

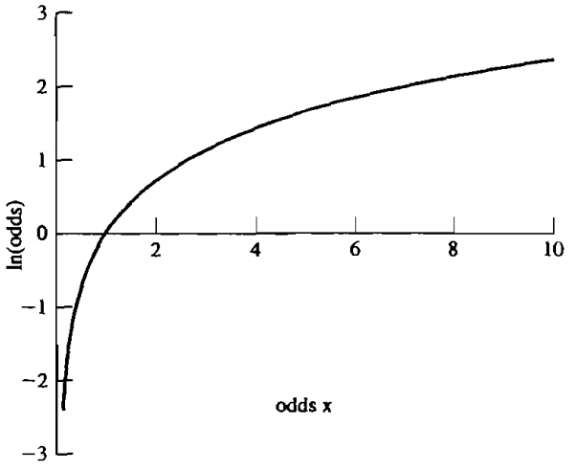
2.5.2 Mô hình Logit

Thay vì mô hình hóa cho hàm xác suất $p(x)$ theo mô hình hồi quy tuyến tính, chúng ta xét tỉ số độ lệch (odds ratio):

$$\text{odds} = \frac{p}{1-p} \tag{2.6}$$

Tỉ số độ lệch là tỷ lệ xác suất thuộc lớp 1 và xác suất thuộc lớp 0, tỷ lệ này có thể lớn hơn 1. Ta có thể mô hình hóa trực tiếp tỉ số trên bởi mô hình hồi quy tuyến tính, tuy nhiên với bài toán phân loại nhị phân thông thường sẽ thực hiện phép lấy logarit tự nhiên để giảm sai số trong quá trình tính toán. Do đó, ta định nghĩa hàm $\text{logit}(p)$ như sau:

$$\text{logit}(p) = \ln(\text{odds}) = \ln(\frac{p}{1-p}) \tag{2.7}$$



Hình 2.2: Logarit tự nhiên của tỉ số độ lệch

Trong mô hình đơn giản nhất, ta có thể cho rằng đường logit như một đường thẳng trong dự đoán biến, do đó:

$$\text{logit}(p) = \ln(\text{odds}) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x \tag{2.8}$$

Ta có thể chuyển từ dạng hàm logit hoặc hàm log odds thành hàm phân phối xác suất p:

$$\theta = \frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x) \text{ hay } p(x) = \frac{1}{1+\exp(\beta_0 + \beta_1 x)} = \tag{2.9}$$

Khi đó phương trình tương đương

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)} \text{ hay } p(x) = \frac{1}{1+\exp(-\beta_0 - \beta_1 x)} \tag{2.10}$$

Biểu thức trên có tên gọi là hàm logistic, nó biểu thị sự thay đổi của p(x) nhanh như thế nào so với x. Tuy nhiên không đơn giản để giải thích mối quan hệ của hàm phi tuyến này như đối với mô hình tuyến tính thông thường.

2.5.3 Phân tích hồi quy logistic

Xem xét mô hình với nhiều biến dự đoán. Cho $(x_{j1}, x_{j2}, \ldots, x_{jr})$ là các giá trị của các yếu tố dự đoán r cho quan sát thứ j. Giống mô hình tuyến tính, ta đặt giá đầu tiên bằng 1 và viết:

$$\mathbf{z}_j = [1, z_{1j}, z_{2j}, \ldots, z_{rj}]^T$$

Với giả định này, ta giả sử rằng biến quan sát Y_j tuân theo phân phối Bernoulli với xác suất thành công phụ thuộc vào giá trị của các biến đồng biến. Khi đó:

$$P(Y_j = y_j) = p(\mathbf{z}_j)^{y_j} \cdot (1 - p(\mathbf{z}_j))^{1-y_j}, \quad \text{với } y_j = 0, 1 \tag{2.11}$$

Vậy nên:

$$E(Y_j) = p(\mathbf{z}_j) \quad \text{và} \quad \text{Var}(Y_j) = p(\mathbf{z}_j)(1 - p(\mathbf{z}_j))$$

Đây không phải là giá trị tuân theo mô hình tuyến tính, mà là logarit tự nhiên của tỷ lệ odds. Cụ thể, ta giả sử mô hình:

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r = \boldsymbol{\beta}^T \mathbf{z} \tag{2.12}$$

với:

$$\boldsymbol{\beta} = \left[\beta_0, \beta_1, \ldots, \beta_r\right]^T \tag{2.13}$$

Ước lượng hợp lý cực đại

Các ước lượng của $\boldsymbol{\beta}$ có thể được tìm bằng phương pháp ước lượng tối đa hợp lý (MLE). Hàm hợp lý L được tính bằng phân phối xác suất của các giá trị quan sát y_j. Do đó:

$$L(\beta_0, \beta_1, \ldots, \beta_r) = \prod_{j=1}^{n} p(\mathbf{z}_j)^{y_j} \left(1 - p(\mathbf{z}_j)\right)^{1-y_j} \tag{2.14}$$

Với mô hình logit, ta có:

$$L(\beta_0, \beta_1, \ldots, \beta_r) = \prod_{j=1}^{n} \frac{e^{y_j(\beta_0+\beta_1 z_{1j}+\cdots+\beta_r z_{rj})}}{1 + e^{\beta_0+\beta_1 z_{1j}+\cdots+\beta_r z_{rj}}} \tag{2.15}$$

Các giá trị của tham số tối đa hóa hàm hợp lý này không thể biểu diễn dưới dạng công thức đóng như trong mô hình hồi quy tuyến tính. Thay vào đó, chúng phải được xác định bằng phương pháp số, bắt đầu từ một giá trị khởi tạo và lặp lại cho đến khi đạt cực đại của hàm hợp lý. Kỹ thuật này được gọi là phương pháp bình phương tối thiểu có trọng số lặp lại.

Ta ký hiệu các ước lượng thu được bằng phương pháp tối đa hợp lý là: $\hat{\boldsymbol{\beta}}$

Khoảng tin cậy cho các tham số

Khi cỡ mẫu đủ lớn, ước lượng $\hat{\boldsymbol{\beta}}$ có phân phối xấp xỉ chuẩn với trung bình là $\boldsymbol{\beta}$ (tức là giá trị thực của các tham số), và ma trận hiệp phương sai xấp xỉ:

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \approx \left[\sum_{j=1}^{n} \hat{p}(\mathbf{z}_j)(1 - \hat{p}(\mathbf{z}_j))\mathbf{z}_j \mathbf{z}_j^T\right]^{-1} \tag{2.16}$$

Căn bậc hai của các phần tử trên đường chéo của ma trận này chính là sai số chuẩn ước lượng (standard errors – SE) cho các hệ số $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r$.

Với cỡ mẫu lớn, khoảng tin cậy 95% cho mỗi tham số β_k được cho bởi:

$$\hat{\beta}_k \pm 1.96 \cdot SE(\hat{\beta}_k), \quad k = 0, 1, \ldots, r \tag{2.17}$$

Khoảng tin cậy cũng giúp đánh giá tầm quan trọng của từng thành phần trong mô hình logit. Với mô hình có r biến dự đoán cộng thêm hằng số, ta ký hiệu giá trị lớn nhất của hàm hợp lý (2.15) là:

$$L_{\max} = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r) \tag{2.18}$$

- Kiểm định tỉ số hợp lý (Likelihood Ratio Test) dùng để kiểm định giả thuyết đơn H_0 : $\beta_k = 0$, ta xét tỉ số độ lệch sau:

$$-2\ln\left(\frac{L_{\text{max, Reduced}}}{L_{\max}}\right) \tag{2.19}$$

trong đó $L_{\text{max, Reduced}} = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{k-1}, \hat{\beta}_{k+1}, \ldots, \hat{\beta}_r)$. Độ lệch (2.19) xấp xỉ $\chi^2(1)$ khi mô hình ít hơn một biến dự báo. Giả thuyết H_0 bị bác bỏ khi độ lệch trên lớn

- Kiểm định Wald (Wald Test) dùng để kiểm định giả thuyết đơn $H_0 : \beta_k = 0$:

$$Z = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \tag{2.20}$$

2.5.4 Mô hình hồi quy logistic với bài toán phân loại nhị phân

Giờ đây chúng ta xem xét một trường hợp tổng quát hơn một chút, trong đó có nhiều lần thử được thực hiện tại cùng các giá trị của biến đồng biến z_j và có tổng cộng m tập hợp khác nhau mà tại đó các biến dự đoán là không đổi. Khi n_j lần thử độc lập được tiến hành với các biến dự đoán là z_j, phản hồi Y_j được mô hình hóa như một phân phối nhị thức với xác suất $p(z_j) = P(\text{Thành công} \mid z_j)$. Vì các Y_j được giả định là độc lập, nên hàm hợp lý là tích:

$$L(\beta_0, \beta_1, \ldots, \beta_r) = \prod_{j=1}^{m} \binom{n_j}{y_j} p(z_j)^{y_j}(1 - p(z_j))^{n_j - y_j} \tag{2.21}$$

trong đó các xác suất $p(z_j)$ tuân theo mô hình logit. Ước lượng hợp lý cực địa $\hat{\beta}$ trong trường hợp này chỉ có thể giải bằng các phương pháp số do không có công thức nghiệm. Khi kích thước mẫu đủ lớn, ma trận hiệp phương sai $\widehat{\text{Cov}}(\hat{\beta})$ có giá trị xấp xỉ là:

$$\widehat{\text{Cov}}(\hat{\beta}) \approx \left[\sum_{j=1}^{n} n_j \hat{p}(\mathbf{z}_j)(1 - \hat{p}(\mathbf{z}_j))\mathbf{z}_j \mathbf{z}_j^T\right]^{-1} \tag{2.22}$$

và vị trí thứ i trong đường chéo chính của ma trận này là một ước lượng điểm cho phương sai $\widehat{\beta}_{i+1}$, đồng thời căn bậc hai của giá trị này cũng là một ước lượng điểm cho độ lệch chuẩn $SE(\hat{\beta}_{i+1})$.

Ta cũng nhận thấy ước lượng điểm giá trị phương sai của xác suất $\hat{p}(x_j)$ được xấp xỉ bởi:

$$\text{Var}(\hat{p}(z_k)) \approx [\hat{p}(z_k)(1 - \hat{p}(z_k))]^2 z_k^T \left[\sum_{j=1}^{m} n_j \hat{p}(z_j)(1 - \hat{p}(z_j)) z_j z_j^T\right]^{-1} z_k. \tag{2.23}$$

Kiểm định mô hình

Trong nhiều trường hợp, ta cần kiểm tra tính phù hợp của từng mô hình đối với dữ liệu qua những câu hỏi sau:

- Có tồn tại sai số có hệ thống (systematic departure), tức là khi kiểm tra với dữ liệu ta thấy các kết quả không chính xác theo cách có hệ thống trong mô hình logistic đã chọn hay không?

- Có tồn tại điểm dữ liệu nào "không bình thường" (outliers) mà không khớp với dạng điệu (pattern) tổng thể của dữ liệu không?
- Có tồn tại điểm dữ liệu nào khi thêm vào hoặc xóa đi sẽ ảnh hưởng lớn (high-influence) đến kết quả của mô hình không?

Nếu không tồn tại mô hình cho $p(z_j) = P(\text{Phép thử thành công} \mid x_j)$, ta có thể ước lượng bằng cách quan sát số lượng phép thử thành công (hay số các số 1) y_j trong n_j phép thử. Xác suất thành công của trường hợp thứ j là:

$$\binom{n_j}{y_j} p^{y_j}(z_j)(1 - p(z_j))^{n_j - y_j}. \tag{2.24}$$

Biểu thức này đạt giá trị cực đại khi $\hat{p}(z_j) = \frac{y_j}{n_j}$, với mọi $j = 1, 2, \ldots, n$. Khi đó $m = \sum n_j$ và:

$$-2 \ln L_{\text{max, phi tham số}} = -2 \sum_{j=1}^{m} \left[y_j \ln \left(\frac{y_j}{n_j} \right) + (n_j - y_j) \ln \left(1 - \frac{y_j}{n_j} \right) \right] + 2 \ln \left(\prod_{j=1}^{m} \binom{n_j}{y_j} \right). \tag{2.25}$$

Ta đồng thời có thể xác định độ lệch giữa mô hình phi tham số và một mô hình đã khớp với một hằng số và $r - 1$ biến dự đoán là:

$$G^2 = 2 \sum_{j=1}^{m} \left[y_j \ln \left(\frac{y_j}{\hat{y}_j} \right) + (n_j - y_j) \ln \left(\frac{n_j - y_j}{n_j - \hat{y}_j} \right) \right] \tag{2.26}$$

trong đó $\hat{y}_j = n_j \hat{p}(z_j)$ là số lần thành công được dự đoán. Đây là độ lệch cụ thể đóng vai trò tương tự như tổng bình phương phần dư trong các mô hình tuyến tính.

Với kích thước mẫu lớn, G^2 xấp xỉ phân phối chi bình phương với số bậc tự do f bằng số lượng quan sát m trừ đi số lượng tham số β được ước lượng.

Lưu ý rằng độ lệch đối với mô hình đầy đủ G^2_{Full}, và độ lệch đối với mô hình rút gọn G^2_{Reduced}, tạo nên một thành phần cho các biến dự đoán bổ sung:

$$G^2_{\text{Reduced}} - G^2_{\text{Full}} = -2 \ln \left(\frac{L_{\text{max, Reduced}}}{L_{\text{max}}} \right) \tag{2.27}$$

Sự khác biệt này xấp xỉ phân phối chi bình phương với số bậc tự do $df = df_{\text{Reduced}} - df_{\text{Full}}$. Một giá trị lớn của sự khác biệt ngụ ý rằng mô hình đầy đủ là cần thiết.

Khi m lớn, có quá nhiều xác suất cần ước lượng trong mô hình phi tham số, và việc xấp xỉ phân phối chi bình phương không thể được thiết lập bởi các phương pháp chứng minh hiện có. Khi đó, tốt hơn nên dựa vào kiểm định tỷ số khả năng hợp lý cho các mô hình logistic khi chỉ loại bỏ một vài biến.

Phần dư và kiểm định độ phù hợp

Ta có thể định nghĩa về phần dư như sau:

- Phần dư độ lệch (Deviance Residuals):

$$d_j = \pm \sqrt{2 \left[y_j \ln \left(\frac{y_j}{n_j \hat{p}(z_j)} \right) + (n_j - y_j) \ln \left(\frac{n_j - y_j}{n_j (1 - \hat{p}(z_j))} \right) \right]} \tag{2.28}$$

  Dấu của d_j giống với dấu của $y_j - n_j \hat{p}(z_j)$, và:

- Nếu $y_j = 0$, thì $d_j = -\sqrt{2n_j|\ln(1-\hat{p}(z_j))|}$
- Nếu $y_j = n_j$, thì $d_j = -\sqrt{2n_j|\ln\hat{p}(z_j)|}$

- Phần dư Pearson(Pearson Residual)

$$r_j = \frac{y_j - n_j\hat{p}(z_j)}{\sqrt{n_j\hat{p}(z_j)(1-\hat{p}(z_j))}} \tag{2.29}$$

- Phần dư Pearson chuẩn hóa (Standardized Pearson residuals) dùng để hiệu chỉnh phương sai không bằng nhau trong phần dư thô bằng cách chia cho độ lệch chuẩn:

$$r_{sj} = \frac{r_j}{\sqrt{1-h_{jj}}} \tag{2.30}$$

Một kiểm định tổng thể cho độ phù hợp, đặc biệt hiệu quả khi kích thước mẫu nhỏ, là thống kê chi-bình phương của Pearson:

$$X^2 = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} \frac{(y_j - n_j\hat{p}(z_j))^2}{n_j\hat{p}(z_j)(1-\hat{p}(z_j))} \tag{2.31}$$

Thống kê chi bình phương là một giá trị tóm tắt cho độ phù hợp, bằng tổng bình phương của các phần dư Pearson. Việc quan sát phần dư Pearson cho phép đánh giá chất lượng mô hình trên toàn bộ miền giá trị của các biến giải thích.

Chương 3

Ứng dụng vào bài toán dự đoán

3.1 Khám phá dữ liệu

Bộ dữ liệu được lấy trên trang Kaggle, thu thập thông tin cung cấp dịch vụ của một viễn thông để dự đoán khách hàng có rời bỏ ngành dịch vụ hay không ?

Thống kê mô tả

Tập dữ liệu gồm 7043 bản ghi, mỗi bản ghi tương ứng với thông tin của một khách hàng đã và đang sử dụng dịch vụ, bao gồm 21 trường dữ liệu được mô tả cụ thể sau đây:

STT	Tên cột	Giải thích
1	customerID	Mã khách hành
2	gender	Giới tính
3	Partner	Khách hàng có sống với vợ/chồng hay không
4	SeniorCitizen	Khách hàng có phải là người cao tuổi không
4	Partner	Khách hàng có người phụ thuộc không
5	tenure	Số tháng khách hàng đã sử dụng dịch vụ
6	PhoneService	Khách hàng có đăng ký dịch vụ điện thoại không
7	MultipleLines	Có sử dụng nhiều đường dây điện thoại không
8	InternetService	Loại dịch vụ internet
9	OnlineSecurity	Có dùng dịch vụ bảo mật internet không.
10	OnlineBackup	Có dùng dịch vụ sao lưu trực tuyến không.
11	DeviceProtection	Có dùng dịch vụ bảo vệ thiết bị không.
12	TechSupport	Có dùng hỗ trợ kỹ thuật không.
13	StreamingTV	Có dùng dịch vụ xem TV trực tuyến không.
14	StreamingMovies	Có dùng dịch vụ xem phim trực tuyến không.
15	Contract	Loại hợp đồng.
16	PaperlessBilling	Có nhận hóa đơn điện tử không
17	PaymentMethod	Phương thức thanh toán
18	MonthlyCharges	Phí dịch vụ hàng tháng.
19	TotalCharges	Tổng số tiền đã thanh toán từ khi đăng ký đến nay.
20	Churn	Khách hàng có rời bỏ không

Ta xem xét một số giá trị thống kê cơ bản của các thuộc tính số bao gồm: trung bình, tứ phân vị, giá trị lớn nhất, giá trị bé nhất,...

	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7032.000000
mean	32.371149	64.761692	2283.300441
std	24.559481	30.090047	2266.771362
min	0.000000	18.250000	18.800000
25%	9.000000	35.500000	401.450000
50%	29.000000	70.350000	1397.475000
75%	55.000000	89.850000	3794.737500
max	72.000000	118.750000	8684.800000

Hình 3.1: Thống kê mô tả dữ liệu số

3.2 Tiền xử lý dữ liệu.

Tiền xử lý dữ liệu là bước đầu tiên trong quy trình xử lý dữ liệu, nhằm chuẩn bị dữ liệu thô để sẵn sàng cho các giai đoạn phân tích và xử lý tiếp theo. Dữ liệu thô thường có thể chứa lỗi, thiếu sót hoặc không đồng nhất, do đó, tiền xử lý dữ liệu giúp cải thiện chất lượng và tính toàn vẹn của dữ liệu. Một số công việc trong tiền xử lý dữ liệu có thể liệt kê ra như sau.

Loại bỏ cột không cần thiết.

Những cột thông tin không cần thiết cho mô hình nên được xóa đi để tránh gây nhiễu và tăng hiệu suất của mô hình lên.

Trong mô hình phân loại, cột "customerID" không có ý nghĩa nên ta sẽ lựa chọn bỏ cột này.

Xử lý dữ liệu bị thiếu.

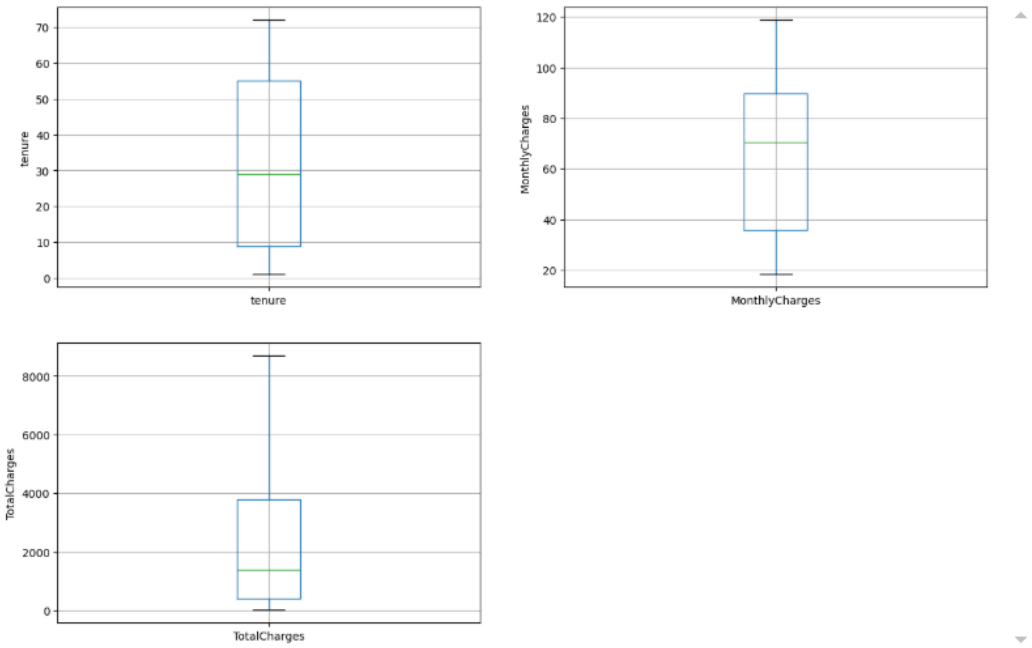
Xử lý dữ liệu thiếu là bước làm sạch dữ liệu bằng cách điền giá trị thay thế hoặc loại bỏ các dòng/cột chứa giá trị bị thiếu, giúp đảm bảo mô hình không bị sai lệch hoặc lỗi khi huấn luyện.

Sau khi kiểm tra thì ta thấy rằng trường thông tin "TotalCharges" có 11 dữ liệu thiếu, chiếm 0.16% toàn bộ dữ liệu. Tỉ lệ của dữ liệu thiếu rất nhỏ nên ta có thể xóa đi những dòng đó mà mô hình không bị ảnh hưởng.

Kiểm tra điểm ngoại lai

Điểm ngoại lai (Outliter) là những giá trị bất thường so với các giá trị khác, có thể gây ra sự sai lệch trong kết quả dự đoán của mô hình.

Ta kiểm tra điểm ngoại lai của các trường định dạng số ("tenure", "MonthlyCharges", "Total Charges") dựa vào biểu đồ Boxplot.

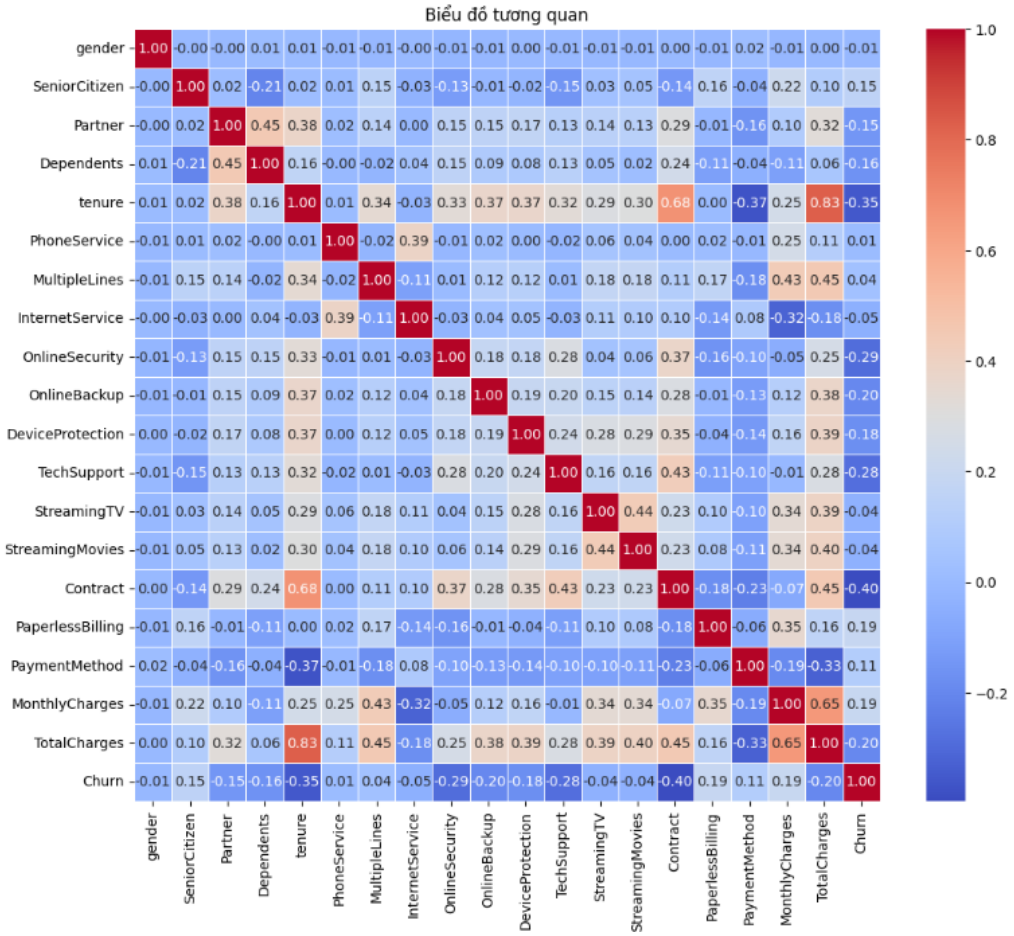


Hình 3.2: Box plot kiểm tra điểm ngoại lai

Ta nhận thấy rằng bộ dữ liệu không có điểm ngoại lai.

3.3 Trực quan hóa dữ liệu

Ma trận tương quan

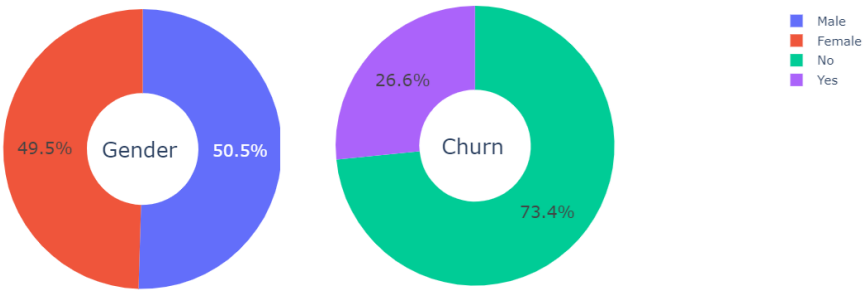


Hình 3.3: Ma trận tương quan

Ma trận tương quan thể hiện mối quan hệ giữa các biến. Hầu như các biến đều không tương quan với nhau (hệ số tương quan nhỏ). Chỉ có hệ số giữa "TotalCharges"-"tenure" có hệ số tương quan cao là 0.83 và giữa "Contract"-"tenure" có hệ số tương quan tương đối là 0.68. Đó là điều dễ hiểu, bởi vì khách hàng sử dụng dịch vụ trong nhiều tháng thì sẽ có tổng số tiền tiêu sẽ cao hơn, tương tự với khách hàng kí hợp đồng dài hạn thì sẽ sử dụng dịch vụ lâu hơn.

Tiếp theo ta sẽ phân tích thông tin của khách hàng để khảo sát xu hướng rời bỏ ngành dịch vụ.

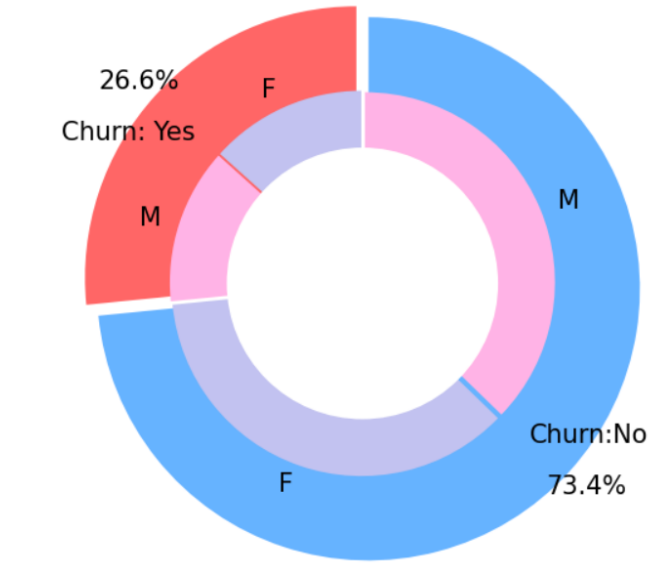
3.3.1 Giới tính (Gender) và tỉ lệ rời bỏ



Hình 3.4: Khảo sát giới tính khách hàng

Theo khảo sát, ta thấy:

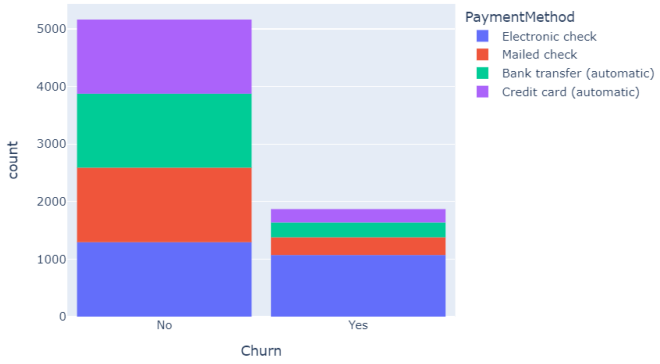
- Có 26.6% khách hàng rời bỏ ngành dịch vụ
- Có 49.5% người tiêu dùng là nữ và đối với nam là 50.5%



Hình 3.5: Tỉ lệ rời bỏ theo giới tính khách hàng

Dựa vào hai biểu đồ trên, ta có thể thấy rằng việc khách hàng rời bỏ ngành dịch vụ không liên quan đến giới tính khách hàng. Cả hai giới tính đều có xu hướng cũng như nhu cầu sử dụng tương đương nhau.

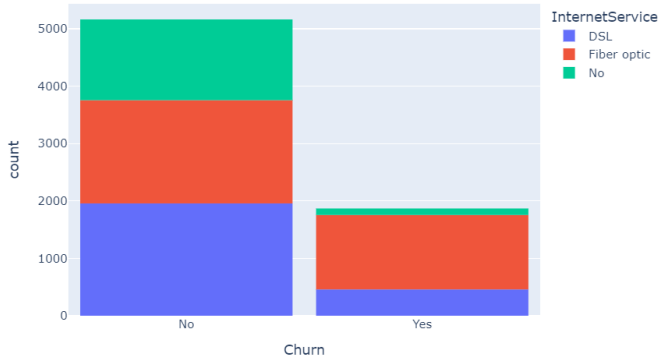
3.3.2 Phương thức thanh toán (Payment Method) và tỉ lệ rời bỏ



Hình 3.6: Tỉ lệ rời bỏ theo phương thức thanh toán

Có thể nhận thấy rằng những người trả bằng séc điện tử (Electronic Check) có xu hướng rời bỏ ngành dịch vụ cao hơn

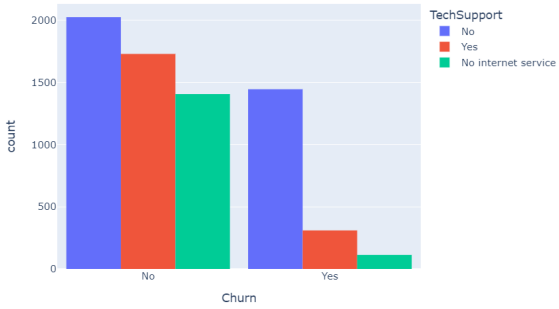
3.3.3 Dịch vụ mạng (Internet Service) và tỉ lệ rời bỏ



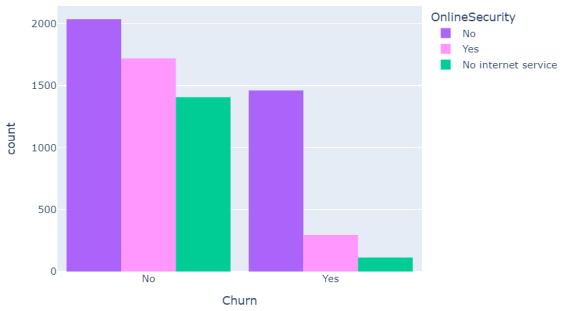
Hình 3.7: Tỉ lệ rời bỏ theo dịch vụ mạng

Khách hàng lựa chọn dịch vụ cáp quang (Fiber optic) có tỉ lệ rời bỏ cao hơn so với các loại dịch vụ mạng khác. Ngoài ra thì khách hàng sử dụng dịch vụ DSL dường như rất hài lòng với dịch vụ của nhà mạng

3.3.4 Hỗ trợ kỹ thuật (Tech support) - Bảo mật trực tiếp (Online security) và tỉ lệ rời bỏ.



(a) Tỉ lệ rời bỏ theo dịch vụ hỗ trợ kĩ thuật



(b) Tỉ lệ rời bỏ theo dịch vụ bảo mật trực tuyến

Hình 3.8: Tỉ lệ rời bỏ theo từng loại dịch vụ hỗ trợ và bảo mật

Một điều rõ ràng rằng những khách hàng không đăng kí bảo mật trực tuyến (Online security) hoặc không có hỗ trợ kỹ thuật (Tech support) sẽ có tỉ lệ rời bỏ cao hơn.

3.4 Xây dựng mô hình bài toán

3.5 Đánh giá mô hình

Kết luận

Tài liệu tham khảo