TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BUSINESS INTELLIGENCE
FINAL PROJECT

DỰ ĐOÁN KHÁCH HÀNG CÓ SỬ DỤNG PHIẾU QUÀ TẶNG TRÊN ỨNG DỤNG THANH TOÁN

Giảng viên hướng dẫn: Thầy DƯƠNG HỮU PHÚC
Sinh viên 1: TRƯƠNG ĐÌNH ÁNH
Sinh viên 2: VÕ NHẬT DUY
Sinh viên 3: NGUYỄN CHÍ KHÂM

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1 Giới thiệu đề tài

Với sự phát triển mạnh mẽ của khoa học kỹ thuật trong thời đại số, cùng với việc ứng dụng công nghệ vào mô hình kinh doanh đã mở ra một kỷ nguyên mới, kỷ nguyên của thương mại thông minh. Với tất cả những tiện ích nó đem lại trong suốt hơn 10 năm qua, nhanh, gọn, tiện lợi, dễ tiếp cận cho cả khách hàng lẫn chủ doanh nghiệp. Việc tận dụng tối đa sức mạnh của Công nghệ thông tin vào lĩnh vực kinh doanh đã là một xu hướng mà bất cứ doanh nghiệp lớn mạnh nào đều hướng đến.

Không chỉ là những lợi ích mà ta dễ dàng nhìn thấy suốt 10 năm qua, các doanh nghiệp đã bắt đầu tìm kiếm, nghiên cứu, mở rộng các giải pháp nhằm hỗ trợ, giúp đỡ, thậm chí là ra quyết định cho mô hình kinh doanh của mình đạt hiệu quả cao. Vì vậy, công nghệ máy học đã được ứng dụng một cách mạnh mẽ vào kinh doanh, từ những dữ liệu của doanh nghiệp, hệ thống sẽ đưa ra gợi ý các chiến lược, định hướng cho tương lai, từ đó giúp doanh nghiệp tìm và giữ được nhiều khách hàng.

Đi từ những vấn đề được nêu trên và nhận thấy tầm quan trọng và sức mạnh của thương mại thông minh, nhóm chúng em sẽ thực hiện đồ án với đề tài "Dự đoán khách hàng sử dụng mã giảm giá trên ứng dụng thanh toán".

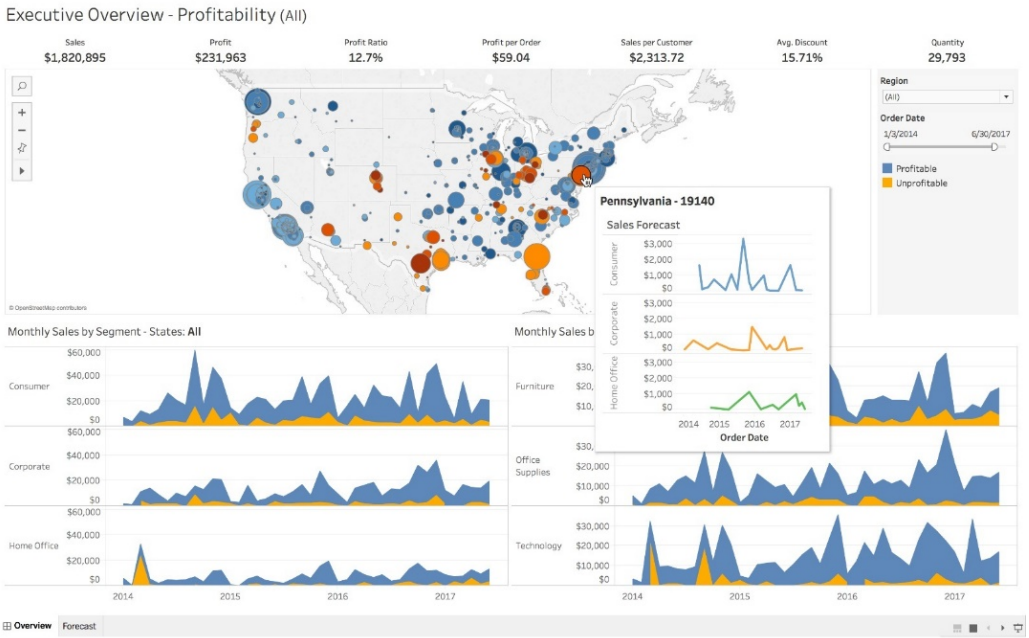
2 Phát biểu bài toán

Giả sử chúng em đang có khách hàng là một doanh nghiệp, với một bộ dataset về mã giảm giá được gửi đến các khách hàng với các tình huống, điều kiện khác nhau. Bộ dataset được cung cấp này chứa các dữ liệu mô tả rất nhiều tình huống khác nhau của một cá nhân, bao gồm các thuộc tính như: thời tiết, địa điểm hay đến, thời gian, hay đi ăn uống cùng ai, thời hạn của mã giảm giá, ... và thuộc tính quan trọng nhất là với các điều kiện trên người đó có quyết định sử dụng mã giảm giá hay không. Nhiệm vụ của chúng em là sẽ giúp doanh nghiệp sử dụng bộ dữ liệu này một cách triệt để.

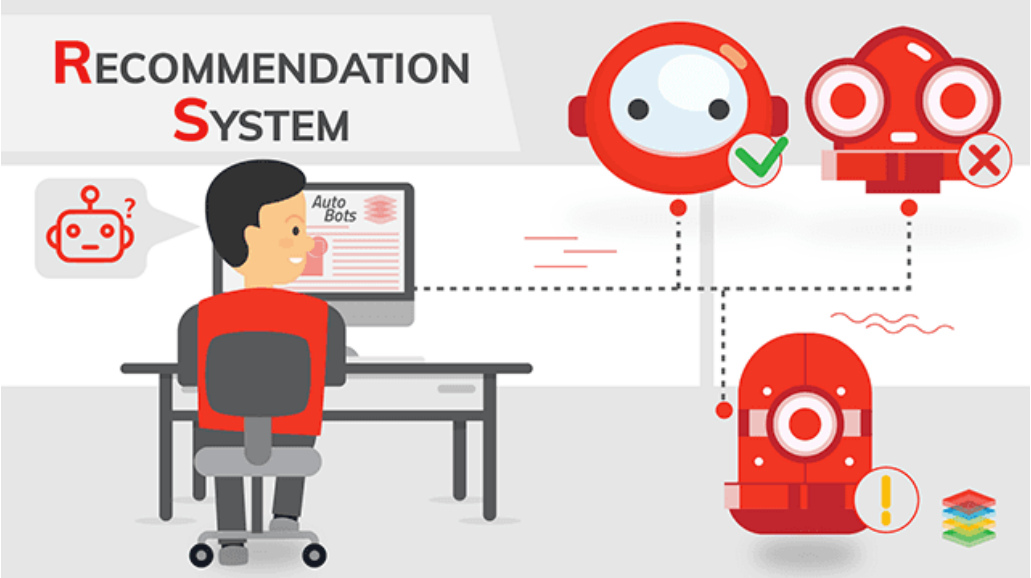
3 Mục tiêu đề tài

Khi đã có sẵn bộ dữ liệu về mã giảm giá được gửi đến các khách hàng của doanh nghiệp, chúng em sẽ thông qua việc trực quan hóa dữ liệu ấy để giúp nó kể nên câu chuyện kinh doanh hiện tại và tương lai, xây dựng một hệ thống gợi ý hiệu quả giúp họ đưa ra các chiến lược và kế hoạch kinh doanh hiệu quả cao trong tương lai. Thông qua những việc trên chúng em đồng thời sẽ học tập được những kiến thức và kỹ năng mới.

Đối với việc trực quan hóa dữ liệu, chúng em sẽ sử dụng phần mềm Tableau, một phần mềm được dùng rất nhiều trong ngành BI (Bussiness Intelligence) chỉ thông qua việc kéo thả. Nó giống như Excel nhưng sẽ thể hiện dữ liệu thành hình ảnh, biểu đồ một cách sinh động và dễ hiểu.



Đối với việc xây dựng hệ thống gợi ý, chúng em sẽ tìm hiểu và nghiên cứu các thuật toán của học máy bao gồm: Decision Tree, Naive Bayes, Random Forest, Logistic Regression. Chúng em sẽ trình bày lý thuyết về nguyên lý của từng thuật toán, áp dụng chúng vào dataset thông qua tính toán và demo. Các mô hình học máy sau khi được traning sẽ tiến hành dự đoán một khách hàng với các tình huống, hoàn cảnh cụ thể sẽ có sử dụng voucher hay không. Cuối cùng so sánh độ hiệu quả giữa chúng và chọn ra thuật toán có hiệu quả nhất.



4 Phạm vi đề tài

Đề tài này xoay quanh lĩnh vực Thương mại thông minh. Đối tượng nghiên cứu là những khách hàng cũ sau đó phân tích hành vi của khách hàng cũ mà ứng dụng mã khuyến mãi cho khách hàng mới.

5 Ý nghĩa khoa học và thực tiễn

Thông qua việc trực quan hóa dữ liệu bằng phần mềm Tableau, chúng em sẽ kể được một câu chuyện mà bộ data ấy muốn nói, về hiện tại và tương lai của doanh nghiệp cần phải làm gì.

Mô hình dự đoán cũng sẽ góp phần giúp đỡ đưa ra các gợi ý thông qua việc dự đoán một khách hàng với ngữ cảnh đó thì họ có dùng voucher hay không nhằm offer cho người phù hợp, thúc đẩy doanh thu.

Bộ dataset này không phải dữ liệu được thu từ thực tế, nó chỉ là một cuộc khảo sát trên Amazon Mechanical Turk, chúng em đang giả lập nó là một dataset của một công ty thu được và giúp họ đinh hướng chiến lược giảm giá. Liên hệ đề tài với thực tiễn, chúng em hoàn toàn có thể thực hiện mở rộng nó với các dự án thực tế với các thuộc tính điều kiện khác hơn và đưa ra gợi ý, một số chiến lược thực tế như: tổ chức chiếc lược phát voucher hiệu quả, dự đoán khách hàng có thích sản phẩm đó hay không, ...

6 Cấu trúc báo cáo

Mục lục

CHƯƠNG 2: TỔNG QUAN GIẢI THUẬT

1 Decision Tree Learning

1.1 Khái niệm

- Decision Tree là cây phân cấp có cấu trúc dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu như là nhị phân(Binary), định danh(Nomial), thứ tự(Ordinal),...
- Hay nói 1 các khác là cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) thì cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

1.2 Mục tiêu ý tưởng của decision trên

- Mục tiêu: tìm một cây nhỏ phù hợp với các training example.
- Ý tưởng: tìm thuộc tính quan trọng nhất để làm gốc mà thuộc tính này có tính đệ quy.

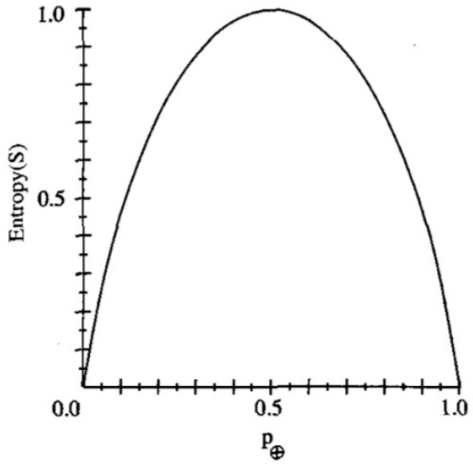
1.3 Giải thuật Decision Tree

Trong giải thuật này thì ta tìm hiểu Entropy và information Gain. Đây là 2 độ đo để chúng ta có thể đi kiếm được thuộc tính có khả năng phân hoạch tốt nhất để chọn làm Node trên cây.

- Entropy: là thước đo tính ngẫu nhiên của dữ liệu đang được xử lý. Entropy càng cao thì khó ra rút ra được kết luật dự đoán.

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

S: là tập dữ liệu trên dataset
c: là classes
Vì sao trong công thức có dấu trừ: vì xác suất miền trị của nó là từ 0->1 mà giá trị nhỏ hơn 1 khi lấy log thì sẽ ra số âm nên thêm dấu trừ vào để khử âm cho ra entropy dương để dễ dàng quan sát.



Tung độ là Entropy(S).
Hoành độ là xác suất của $P_{\oplus}$
⇒ Xác suất của $P_{\oplus}$ nếu rơi về 2 cực 0 hoặc 1 thì kết quả Entropy thấp nhất bằng 0. Khi P = 1 là kết quả dataset đó chỉ trả về 1 kết quả là True còn khi P = 0 là dataset đó trả về 1 kết quả là False. Khi P = 0.5 là Entropy sẽ đạt đỉnh cao nhất.

- Information Gain: là độ lợi của thông tin

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

A: attribute(thuộc tính).
Gain(S,A): là độ lợi thông tin của thuộc tính A trong dataset S.
S_v: là dataset với v nằm trong thuộc tính A.

1.4 Áp dụng giải thuật DTL vào dataset của project cuối kỳ

Ta sẽ chọn ngẫu nhiên 15 dòng trên dataset để áp dụng giải thuật để xử lý dữ liệu dataset và thêm 2 dòng để dự đoán xem khách hàng có quyết định sử dụng dịch vụ của công ty hay không.

destination	passanger	weather	temperature	time	coupon	expiration	gender	age	maritalStatus	has_children	education	occupation
No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	2PM	Restaurant(<20)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	6PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	6PM	Restaurant(<20)	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	55	2PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	55	2PM	Carry out & Take away	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Alone	Sunny	55	10AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	55	6PM	Bar	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	55	6PM	Restaurant(20-50)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	80	6PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Bar	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	80	7AM	Restaurant(20-50)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	80	7AM	Carry out & Take away	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Restaurant(<20)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering

income	Bar	CoffeeHouse	CarryAway	RestaurantLessThan20	Restaurant20To50	toCoupon_GEQ5min	toCoupon_GEQ15min	toCoupon_GEQ25min	direction_same	direction_opp	Y
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	1	0	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	1	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	1	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	1	0	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	

Đầu tiên ta tìm Entropy của cột Y
Entropy([7+,8-]) = 0.997
Sau đó ta xét từng thuộc tính trên dataset và tìm giá trị Gain của thuộc tính Y với từng thuộc tính có trong dataset

+ Values(destionation) = No Urgent Place, home, work
  S(No Urgent Place) = [3+,4-] ⇒ E(nup) = 0.985
  S(home) = [1+,2-] ⇒ E(home) = 0.918
  S(work) = [3+,2-] ⇒ E(work) = 0.97
  ⇒ Gain(S, destination) = 0.03

+ Values(passanger) = friend(s), alone
  S(friend(s)) = [3+,3-] ⇒ E(friend) = 1
  S(alone) = [4+,5-] ⇒ E[alone] = 0.991
  ⇒ Gain(S, passanger) = 0.002

+ Values(weather) = sunny
  S(sunny) = [7+,8-] ⇒ E(sunny) = 0.997
  ⇒ Gain(S, weather) = 0

+ Values(temparature) = 55, 80
  S(55) = 4+,4-] ⇒ E(55) = 1
  S(80) = [3+,4-] ⇒ E(80) = 0.985
  ⇒ Gain(S, temperature) = 0.004

+ Values(time) = 2PM, 6PM, 7AM, 10AM
  S(2PM) = [2+,2-] ⇒ E(2PM) = 1
  S(6PM) = [2+,3-] ⇒ E(6PM) = 0.97
  S(7AM) = [3+2-] ⇒ E(7AM) = 0.97
  S(10AM) = [0+1-] ⇒ E(10AM) = 0
  ⇒ Gain(S, time) = 0.084

+ Values(coupon) = Restaurant(<20), Restaurant(20-50), Coffee House, Carry out & Take away, Bar
  S(Restaurant(<20)) = [3+,0-] ⇒ E(Restaurant(<20)) = 0
  S(Restaurant(20-50)) = [0+,2] ⇒ E(Restaurant(20-50)) = 0
  S(Coffee House) = [0+,6-] ⇒ E(Coffee House) = 0
  S(Carry out & Take away) = [2+,0-] ⇒ E(Carry out & Take away) = 0
  S(Bar) = [2+,0-] ⇒ E(Bar) = 0
  ⇒ Gain(S, coupon) = 0.997

+ Values(expiration) = 1d, 2h
  S(1d) = [5+,3-] ⇒ E(1d) = 0.954
  S(2h) = [2+,5-] ⇒ E(2h) = 0.863
  ⇒ Gain(S, expiration) = 0.0024

+ Values(gender) = Male
  S(Male) = [7+,8-] ⇒ E(Male) = 0.997
  ⇒ Gian(S, gender) = 0

+ Values(age) = 21
  S(21) = [7+,8-] ⇒ E(21) = 0.997
  ⇒ Gain(S, age) = 0

+ Values(maritalStatus) = Single
  S(Single) = [7+,8-] ⇒ E(Single) = 0.997
  ⇒ Gain(S, maritalStatus) = 0

+ Values(Has_children) = 0
  S(0) = [7+,8-] ⇒ E(0) = 0.997
  ⇒ Gain(S, Has_children) = 0

+ Values(education) = Bachelors degree
  S(Bachelors degree) = [7+,8-] ⇒ E(Bachelors degree) = 0.997

⇒ Gain(S, education) = 0

+ Values(occupation) = Architecture&Engineering
S(Architecture&Engineering) = [7+,8-] ⇒ E(Student)= 0.997
⇒ Gain(S, occupation) = 0

+ Values(income) = 62500−74999
S(62500−74999) = [7+,8-] ⇒ E(12500−24999) = 0.997
⇒ Gain(S, 62500−74999) = 0

+ Values(Bar) = never
⇒ Gain(S, Bar) = 0

+ Values(CoffeeHouse) = less1
⇒ Gain(S, CoffeeHouse) = 0

+ Values(CarryaAway) = 4~8 ⇒ Gain(S, CarryaAway) = 0

+ Values(RestaurantLessThan20) = 4
~
8
⇒ Gain(RestaurantLessThan20) = 0

+ Values(Restaurant20To50) = less1
⇒ Gain(Restaurant20To50) = 0

+ Values(toCoupon_GEQ5min) = 1
⇒ Gain(toCoupon_GEQ5min) = 0

+ Values(toCoupon_GEQ15min) = 0, 1
⇒ S(0) = [3+,4-] ⇒E(0) = 0.985
S(1) = [4+,4] ⇒ E(1) = 1
⇒ Gain(S, toCoupon_GEQ15min) = 0.004

+ Values(toCoupon_GEQ25min) = 0, 1
S(0) = [6+,7-] ⇒ E(0) = 0.996
S(1) = [1+,1] ⇒ E(1) = 1
⇒ Gain(S, toCoupon_GEQ25min) = 0.0004

+ Values(direction_same) = 0, 1
S(0) = [5+,8-] ⇒ E(0) = 0.961
S(1) = [2+,0] ⇒ E(1) = 0
⇒ Gain(S, direction_same) = 0.164

+ Values(direction_opp) = 0, 1
S(0) = [2+,0-] ⇒ E(0) = 0
S(1) = [5+,8-] ⇒ E(1) = 0.961
⇒ Gain(S, direction_opp) = 0.164

Vậy ta chọn coupon có Gain(S, coupon) cao nhất để làm nút gốc
Coupon có các values là : Restaurant(<20), Restaurant(20-50), Coffee House, Carry out & Take away, Bar
Vì:

S(Restaurant(<20)) = [3+,0-] ⇒ cột Y sẽ trả về 1

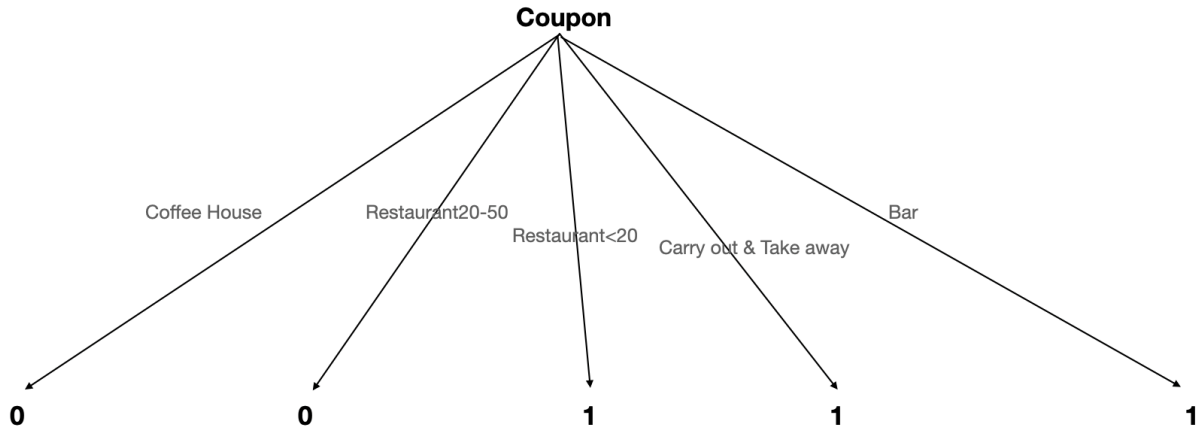
S(Restaurant(20-50)) = [0+,2] ⇒ cột Y sẽ trả về 0

S(Coffee House) = [0+,6-] ⇒ cột Y sẽ trả về 0

S(Carry out & Take away) = [2+,0-] ⇒ cột Y sẻ trả về 1

S(Bar) = [2+,0-] ⇒ cột Y sẽ trả về 1

Do lựa chọn ngẫu nhiên 15 dòng trên dataset nên đã xảy ra trường hợp đặc biệt như hình vẽ ở dưới



2 Random Forest Learning

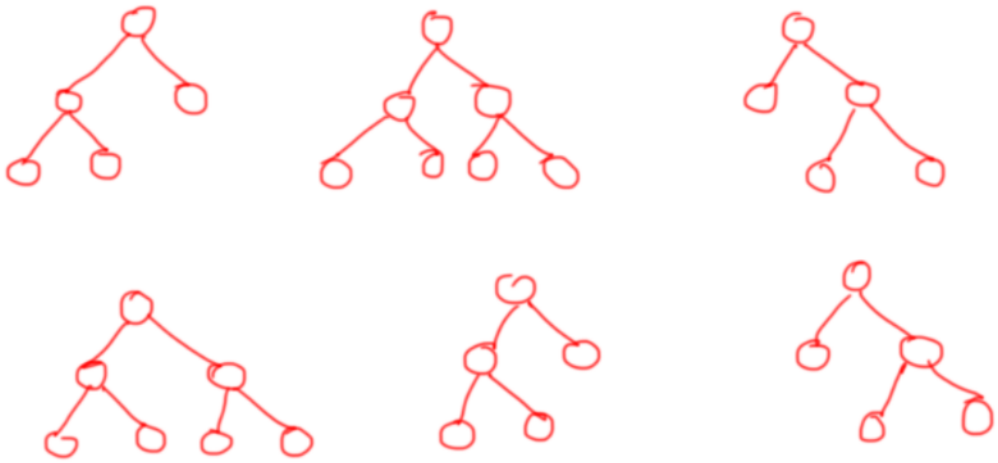
2.1 Khái niệm:

Random forest là thuật toán thuộc nhóm supervised learning, có thể sử dụng cho bài toán phân lớp và bài toán hồi quy. Thuật toán này sử dụng các cây để làm nền tảng. Random forest là tập hợp của các Decision Tree, mà mỗi cây trong tập hợp đó được chọn theo 1 cách ngẫu nhiên.

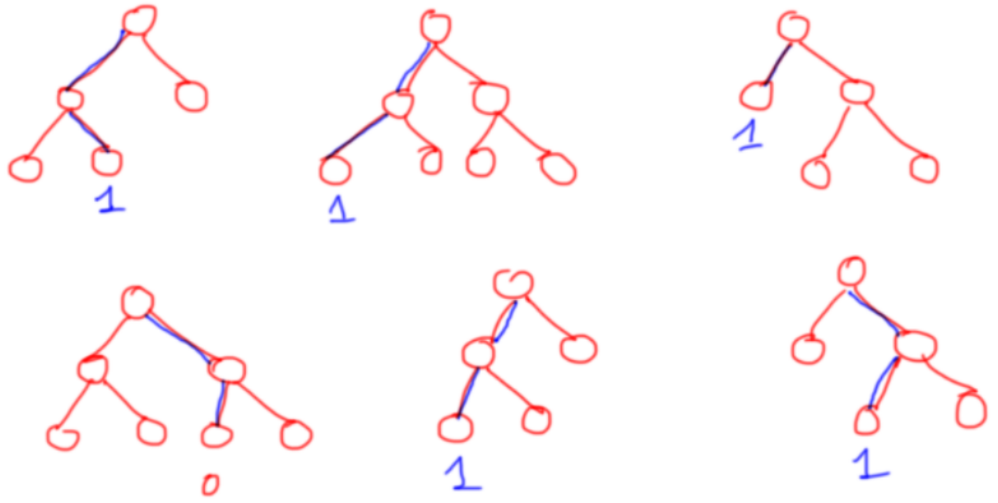
2.2 Tổng quan về thuật toán Random forest:

Random forest nếu dịch ra có nghĩa là rừng ngẫu nhiên, nên thuật toán này sẽ xây dựng nhiều cây quyết định thông qua thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau do có yếu tố ngẫu nhiên ở đây. Kết quả dự đoán cuối cùng sẽ là kết quả được dự đoán nhiều nhất bởi tập hợp cây quyết định đã được xây dựng.

Giả sử chúng ta xây dựng được một tập hợp các cây quyết định một cách ngẫu nhiên.



Sau đó tới bước dự đoán, ta sẽ đưa dữ liệu cần dự đoán vào các cây quyết định đã xây dựng, ở mỗi cây dữ liệu sẽ đi từ trên xuống theo các node điều kiện để trả về các dự đoán. Kết quả dự đoán cuối cùng của thuật toán sẽ là kết quả được trả về nhiều nhất.



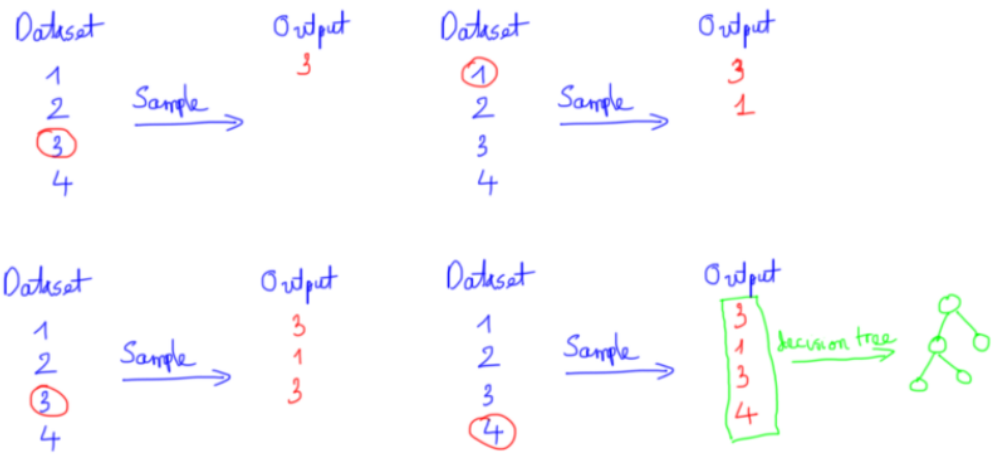
Ở đây kết quả chúng ta có 6 cây quyết định, khi đưa 1 dữ liệu cần dự đoán vào, ta có kết quả 5 cây trả về 1 và 1 cây trả về 0 → Kết quả dự đoán của thuật toán Random forest cho dữ liệu đó sẽ là 1.

2.3 Xây dựng thuật toán Random forest:

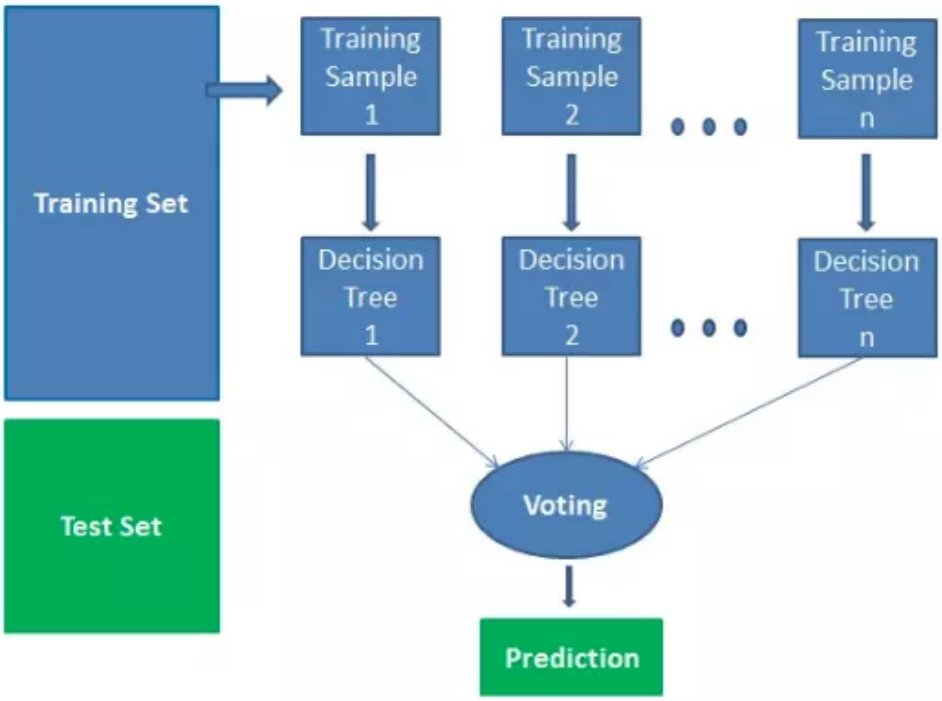
Giả sử ta có bộ dữ liệu với n dữ liệu và d thuộc tính.

Để xây dựng 1 cây quyết định ta sẽ có các bước sau:

1. Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu.
2. Chọn ngẫu nhiên k thuộc tính (k < n), giờ ta đã có n dữ liệu và k thuộc tính mới.
3. Dùng thuật toán Decision tree để xây dựng cây quyết định.
4. Lặp lại bước 1 – 3 cho đến khi ta tìm được số cây quyết định mong muốn.



Do ta xây dựng ra các cây quyết định khác nhau (ngẫu nhiên) nên lúc này các kết quả dự đoán của từng cây có thể sẽ khác nhau. Kết quả dự đoán sẽ được tổng hợp từ các cây quyết định đã xây dựng.



Khi dùng thuật toán Random Forest ta cần chú ý một số thứ sau đây: số lượng cây quyết định mong muốn xây dựng, số lượng thuộc tính để xây dựng cây quyết định. Chi tiết hơn còn có một số yếu tố của cây quyết định như: độ sâu tối đa, số phẩn tử tối thiểu trong 1 node.

2.4 Ưu và nhược điểm:

Ưu điểm

- Có thể được sử dụng cho cả phân loại và hồi quy.
- Dự đoán chính xác.
- Tránh được trường hợp overfitting. Ở thuật toán Decision Tree, nếu ta để một độ sâu đủ lớn, thì cây sẽ phân loại đúng hết các dữ liệu trong tập training dẫn đến mô hình huấn luyện sẽ mắc phải trường hợp overfitting dẫn đến kết quả dự đoán kém chính xác.

– Khắc phục được trường hợp underfitting. Ở thuật toán Random Forest, các cây quyết định được xây dựng ngẫu nhiên từ 2 yếu tố: n dữ liệu ngẫu nhiên, k thuộc tính ngẫu nhiên. Do vậy, thuật toán không dùng tất cả dữ liệu trong tập training, cũng không dùng tất cả thuộc tính nên có thể xảy ra trường hợp underfitting. Để khắc phục điều này, kết quả dự đoán cuối cùng sẽ được bỏ phiếu từ các cây quyết định. Từ đó, các thông tin của cây sẽ bổ sung cho nhau dẫn đến dự đoán kết quả khách quan, khắc phục underfitting

Nhược điểm

– Random forest xử lý chậm.
– Mô hình khó hiểu hơn so với cây quyết định (chỉ cần đưa ra quyết định theo đường dẫn trong cây)

3 Naive Bayes Learning

3.1 Khái niệm

Naive Bayes là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes, thuật toán này thuộc nhóm học có giám sát.//

3.2 Lý thuyết Bayes:

Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện h khi biết D như sau:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(h): xác xuất tiên nghiệm của giả thuyết h
P(D): xác xuất tiên nghiệm của tập dữ liệu training
P(h|D): xác xuất của h khi ta có trước D
P(D|h): xác xuất của D khi ta có trước h

Để tìm được P(h|D) lớn nhất, ta cần tìm được cực đại của P(D|h), ta có công thức:

- Maximum Likelihood hypothesis

$$\begin{aligned} h_{map} &= \underset{h \in H}{\operatorname{argmax}} P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h) \\ &= \underset{h \in H}{\operatorname{argmax}} P(D|h) = h_{ML} \end{aligned}$$

P(h) is uniform distribution over H
It means that P(h_i) = P(h_j) for all h_i and h_i in H

3.3 Giới thiệu thuật toán

- Một tập dữ liệu để học đã được cung cấp sẵn nhằm giúp ta tìm ra được hàm mục tiêu dùng để tìm (dự đoán) giá trị target value. Instance x cần dự đoán sẽ được mô tả bởi các thuộc tính trong tập dữ liệu đã học.
- Mục tiêu: Tìm (dự đoán) giá trị target value hoặc phân loại cho instance x.

3.4 Phân loại Naïve Bayes

- Cách tiếp cận Bayesian là ta sẽ dự đoán giá trị target value có xác xuất cao nhất nhằm phân loại cho instace mới với các thuộc tính $(a_1, a_2, \ldots, a_n)$ mô tả instance có sẵn.

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, \ldots, a_n)$$

$$= \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, \ldots, a_n | v_j) P(v_j)$$

v_j : các target value có thể xảy ra
$(a_1, a_2, \ldots, a_n)$: các thuộc tính

Phân loại Naïve Bayes dựa trên giả thiết đơn giản hóa các giá trị thuộc tính đều là các điều kiện độc lập có điều kiện của các instance và trả các giá trị target value tương ứng. Giả thiết của chúng ta là sẽ cung cấp giá trị target value cho các instance, ta có xác xuất kết hợp của các thuộc tính sẽ chỉ là tích của tất cả các xác suất cho từng thuộc tính riêng biệt

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, \ldots, a_n | v_j) P(v_j)$$

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

3.5 Kỹ thuật làm mịn:

3.5.1 Vấn đề:

Khi ta tính toán xác suất trả về các target value của từng giá trị của các thuộc tính.

Ví dụ:

$$P(Wind = Strong | PlayTennis = N) = \frac{n_C}{n} = \frac{3}{5}$$

n: là tổng số PlayTennis là N trong tập training.
n_c: là số Wind = Strong trong tổng số PlayTennis là N

Điều này đôi khi sẽ dẫn đến xảy ra 2 khó khăn cho chúng ta:

- Xác suất rất nhỏ dẫn đến xác suất hướng đến 0 dẫn đến mất đi tính tổng quát ảnh hưởng kết quả dự đoán.
- Xác suất có thể là 0, điều này sẽ dẫn đến kết quả dự đoán sai lệch hoàn toàn (mất đi tính tổng quát)

3.5.2 Giải pháp:

Để tránh những khó khăn đó, ta sẽ làm mịn bằng phương pháp m-estimate:

$$\frac{n_c + mp}{n + m}$$

p: phân phối chuẩn p=$\frac{1}{k}$
m: là hằng số được ta tự định nghĩa dựa vào kinh nghiệm
k: tổng số loại của thuộc tính

3.6 Áp dụng giải thuật Naive Bayes Learning vào dataset của project cuối kì

Ta sẽ chọn ngẫu nhiên 15 dòng trên dataset để áp dụng giải thuật để xử lý dữ liệu dataset và thêm 2 dòng để dự đoán xem khách hàng có quyết định sử dụng dịch vụ của công ty hay không.

destination	passanger	weather	temperature	time	coupon	expiration	gender	age	maritalStatus	has_children	education	occupation
No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	2PM	Restaurant(<20)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	6PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	6PM	Restaurant(<20)	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	55	2PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	55	2PM	Carry out & Take away	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Alone	Sunny	55	10AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	55	6PM	Bar	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	55	6PM	Restaurant(20-50)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Home	Alone	Sunny	80	6PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Bar	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	80	7AM	Restaurant(20-50)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	80	7AM	Carry out & Take away	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Restaurant(<20)	1d	Male	21	Single	0	Bachelors degree	Architecture & Engineering
Work	Alone	Sunny	55	7AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering

income	Bar	CoffeeHouse	CarryAway	RestaurantLessThan20	Restaurant20To50	toCoupon_GEQ5min	toCoupon_GEQ15min	toCoupon_GEQ25min	direction_same	direction_opp	Y
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	1	0	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	1	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	1	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	0
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	1	0	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	0	0	0	1	1
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1	

Ta tạm thời xem 1 là yes, 0 là no.

P(v = Y): $\frac{7}{15}$

P(v = N): $\frac{8}{15}$

Destination	Y	N
No Urgent Place	$\frac{3}{7}$	$\frac{4}{8}$
Home	$\frac{1}{7}$	$\frac{2}{8}$
Work	$\frac{3}{7}$	$\frac{2}{8}$

passenger	Y	N
Friends	$\frac{3}{7}$	$\frac{3}{8}$
Alone	$\frac{4}{7}$	$\frac{5}{8}$

weather	Y	N
sunny	$\frac{7}{7}$	$\frac{8}{8}$

temperature	Y	N
80	$\frac{3}{7}$	$\frac{4}{8}$
55	$\frac{4}{7}$	$\frac{4}{8}$

Time	Y	N
2PM	$\frac{2}{7}$	$\frac{2}{8}$
6PM	$\frac{2}{7}$	$\frac{3}{8}$
10AM	$\frac{0}{7}$	$\frac{1}{8}$
7AM	$\frac{3}{7}$	$\frac{2}{8}$

Coupon	Y	N
Coffee House	0	$\frac{6}{8}$
Restaurant(<20)	$\frac{3}{7}$	0
Carry out & Take away	$\frac{2}{7}$	0
Bar	$\frac{2}{7}$	0
Restaurant(20-50)	0	$\frac{2}{8}$

expiration	Y	N
1d	$\frac{5}{7}$	$\frac{3}{8}$
2h	$\frac{2}{7}$	$\frac{5}{8}$

gender	Y	N
Male	$\frac{7}{7}$	$\frac{8}{8}$

age	Y	N
21	$\frac{7}{7}$	$\frac{8}{8}$

maritalStatus	Y	N
Single	$\frac{7}{7}$	$\frac{8}{8}$

has_children	Y	N
0	$\frac{7}{7}$	$\frac{8}{8}$

education	Y	N
Bachelors degree	$\frac{7}{7}$	$\frac{8}{8}$

occupation	Y	N
Architecture & Engineering	$\frac{7}{7}$	$\frac{8}{8}$

income	Y	N
$62500 - $74999	$\frac{7}{7}$	$\frac{8}{8}$

Bar	Y	N
never	$\frac{7}{7}$	$\frac{8}{8}$

CoffeeHouse	Y	N
less1	$\frac{7}{7}$	$\frac{8}{8}$

CarryAway	Y	N
4~8	$\frac{7}{7}$	$\frac{8}{8}$

RestaurantLessThan20	Y	N
4~8	$\frac{7}{7}$	$\frac{8}{8}$

Restaurant20To50	Y	N
less1	$\frac{7}{7}$	$\frac{8}{8}$

toCoupon_GEQ5min	Y	N
1	$\frac{7}{7}$	$\frac{8}{8}$

toCoupon_GEQ15min	Y	N
1	$\frac{4}{7}$	$\frac{4}{8}$
0	$\frac{3}{7}$	$\frac{4}{8}$

toCoupon_GEQ25min	Y	N
1	$\frac{1}{7}$	$\frac{1}{8}$
0	$\frac{6}{7}$	$\frac{7}{8}$

direction_same	Y	N
1	$\frac{2}{7}$	0
0	$\frac{5}{7}$	$\frac{8}{8}$

direction_opp	Y	N
1	$\frac{5}{7}$	$\frac{8}{8}$
0	$\frac{2}{7}$	0

Vì xác suất độc lập của ta có kết quả bằng 0 nên ta sẽ tiến hành phương pháp làm mịn. Ta chọn m = 6. Ta sẽ được kết quả sau:

p=$\frac{1}{3}$

Destination	Y	N
No Urgent Place	$\frac{5}{13}$	$\frac{3}{7}$
Home	$\frac{3}{13}$	$\frac{2}{7}$
Work	$\frac{5}{13}$	$\frac{2}{7}$

p=$\frac{1}{2}$

passenger	Y	N
Friends	$\frac{6}{13}$	$\frac{3}{7}$
Alone	$\frac{7}{13}$	$\frac{4}{7}$

p=1

weather	Y	N
sunny	1	1

p=$\frac{1}{2}$

temperature	Y	N
80	$\frac{6}{13}$	$\frac{1}{2}$
55	$\frac{7}{13}$	$\frac{1}{2}$

p=$\frac{1}{4}$

Time	Y	N
2PM	$\frac{7}{26}$	$\frac{1}{4}$
6PM	$\frac{7}{26}$	$\frac{9}{28}$
10AM	$\frac{1}{4}$	$\frac{5}{28}$
7AM	$\frac{9}{26}$	$\frac{1}{4}$

p=$\frac{1}{5}$

Coupon	Y	N
Coffee House	$\frac{1}{5}$	$\frac{18}{35}$
Restaurant(<20)	$\frac{21}{65}$	$\frac{1}{5}$
Carry out & Take away	$\frac{16}{65}$	$\frac{1}{5}$
Bar	$\frac{16}{65}$	$\frac{1}{5}$
Restaurant(20-50)	0	$\frac{2}{35}$

p=$\frac{1}{2}$

expiration	Y	N
1d	$\frac{8}{13}$	$\frac{3}{7}$
2h	$\frac{5}{13}$	$\frac{4}{7}$

p=1

gender	Y	N
Male	1	1

p=1

age	Y	N
21	1	1

p=$\frac{1}{1}$

maritalStatus	Y	N
Single	1	1

p=1

has_children	Y	N
0	1	1

p=1

education	Y	N
Bachelors degree	1	1

p=1

occupation	Y	N
Architecture & Engineering	1	1

p=1

income	Y	N
$62500 - $74999	1	1

p=1

Bar	Y	N
never	1	1

p=1

CoffeeHouse	Y	N
less1	1	1

p=1

CarryAway	Y	N
4 ~ 8	1	1

p=1

RestaurantLessThan20	Y	N
4 ~ 8	1	1

p=1

Restaurant20To50	Y	N
less1	1	1

p=1

toCoupon_GEQ5min	Y	N
1	1	1

p=$\frac{1}{2}$

toCoupon_GEQ15min	Y	N
1	$\frac{7}{13}$	$\frac{1}{2}$
0	$\frac{6}{13}$	$\frac{1}{2}$

p=$\frac{1}{3}$

toCoupon_GEQ25min	Y	N
1	$\frac{4}{13}$	$\frac{2}{7}$
0	$\frac{9}{13}$	$\frac{5}{7}$

p=$\frac{1}{2}$

direction_same	Y	N
1	$\frac{5}{13}$	$\frac{1}{2}$
0	$\frac{8}{13}$	$\frac{11}{14}$

p=$\frac{1}{2}$

direction_opp	Y	N
1	$\frac{8}{13}$	$\frac{11}{14}$
0	$\frac{5}{13}$	$\frac{1}{2}$

Dự đoán cho mẫu

Work	Alone	Sunny	55	7AM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1		

P(yes)*P(work|yes)*P(alone|yes)* P(sunny|yes)* P(55|yes)* P(7AM|yes)* P(Coffee House|yes)* P(2h|yes)* P(Male|yes)* P(21|yes)* P(Single|yes)* P(has_children=0|yes)* P(Bachelor degree|yes)* P(Architecture & Engineering|yes)* P($62500 - $74999|yes)* P(never|yes)* P(less1|yes)* P(CarryAway=4~8|yes)* P(RestaurantLessThan20=4~8|yes)* P(Restaurant20to50=less1|yes)* P(toCoupon_GEQ5min=1|yes)* P(toCoupon_GEQ15min=1|yes)* P(toCoupon_GEQ25min=0|yes)* P(direction_same

P(direction_opp=1|yes) =0.0001956

P(yes)*P(work|no)*P(alone|no)* P(sunny|no)* P(55|no)* P(7AM|no)* P(Coffee House|no)* P(2h|no)* P(Male|no)* P(21|no)* P(Single|no)* P(has_children=0|no)* P(Bachelor degree|no)* P(Architecture & Engineering|no)* P($62500 - $74999|no)* P(never|no)* P(less1|no)* P(CarryAway=4~8|no)* P(RestaurantLessThan20=4~8|no)* P(Restaurant20to50=less1|no)* P(toCoupon_GEQ5min=1|no)* P(toCoupon_GEQ15min=1|no)* P(toCoupon_GEQ25min=0|no)* P(direction_same=0|no)* P(direction_opp=1|no) = 0.000705

⇒ Mẫu này ta dự đoán là no

Dự đoán cho mẫu

No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	2h	Male	21	Single	0	Bachelors degree	Architecture & Engineering
$62500 - $74999	never	less1	4~8	4~8	less1	1	1	0	0	1		

P(yes)*P(No Urgent Place|yes)*P(Friends|yes)* P(sunny|yes)* P(80|yes)* P(2pm|yes)* P(CoffeeHouse|yes)* P(2h|yes)* P(Male|yes)* P(21|yes)* P(Single|yes)* P(has_children=0|yes)* P(Bachelor degree|yes)* P(Architecture & Engineering|yes)* P($62500 - $74999|yes)* P(never|yes)* P(less1|yes)* P(CarryAway=4~8|yes)* P(RestaurantLessThan20=4~8|yes)* P(Restaurant20to50=less1|yes)* P(toCoupon_GEQ5min=1|yes)* P(toCoupon_GEQ15min=1|yes)* P(toCoupon_GEQ25min=0|yes)* P(direction_same=0|yes)* P(direction_opp=1|yes)= 0.000117

P(no)*P(No Urgent Place|no)*P(Friends|no)* P(sunny|no)* P(80|no)* P(2pm|no)* P(Coffee House|no)* P(2h|no)* P(Male|no)* P(21|no)* P(Single|no)* P(has_children=0|no)* P(Bachelor degree|no)* P(Architecture & Engineering|no)* P($62500 - $74999|no)* P(never|no)* P(less1|no)* P(CarryAway=4~8|no)* P(RestaurantLessThan20=4~8|no)* P(Restaurant20to50=less1|no)* P(toCoupon_GEQ5min=1|no)* P(toCoupon_GEQ15min=1|no)* P(toCoupon_GEQ25min=0|no)* P(direction_same=0|no)* P(direction_opp=1|no) =0.000793

⇒ Mẫu này ta dự đoán là no

4 Logistic Regression Learning

4.1 Khái niệm

Logistic Regression là thuật toán thuộc nhóm các thuật toán phân lớp. Không giống như hồi quy tuyến tính, thuật toán hồi quy logistic sử dụng hàm sigmiod logistic để trả về một giá trị xác suất có thể được ánh xạ tới hai hay nhiều lớp rời rạc.
Hồi quy tuyến tính khác với hồi quy logistic:

- Hồi quy tuyến tính có thể giúp chúng ta dự đoán được các giá trị liên tục.
- Hồi quy logistic là rời rạc chỉ cho phép các giá trị cụ thể.

destination	passanger	weather	temp	time	coupon	expiratio	gender	age	maritals	has_	educati	occupatior	income	Bar	Coffee	Carr	Resta	Restau	toC	toCo	toCo	direct	direction_	Y
No Urgent Place	Friend(s)	Sunny	80	2PM	Coffee House	1d	Male	21	Single	0	Bachelc	Architecture	$62500 - $	never	less1	4~8	4~8	less1	1	1	0	0	1	0
No Urgent Place	Friend(s)	Sunny	80	2PM	Restaurant(<20)	1d	Male	21	Single	0	Bachelc	Architecture	$62500 - $	never	less1	4~8	4~8	less1	1	1	0	0	1	1
No Urgent Place	Friend(s)	Sunny	80	6PM	Coffee House	2h	Male	21	Single	0	Bachelc	Architecture	$62500 - $	never	less1	4~8	4~8	less1	1	0	0	0	1	0
No Urgent Place	Friend(s)	Sunny	80	6PM	Restaurant(<20)	2h	Male	21	Single	0	Bachelc	Architecture	$62500 - $	never	less1	4~8	4~8	less1	1	1	0	0	1	1

ví dụ:

Dữ liệu về kết quả của khách hàng và mục tiêu của chúng ta là dự đoán liệu khách hàng có sử dụng mã giảm giá ghé vào cửa hàng nào đó khi trên đường di chuyển đến điểm đích hay không dựa trên các đặc trưng của khách hàng đó.

Chúng ta có 24 đặc trưng và giá trị dự đoán như: 1 là chấp nhận và 0 là không chấp nhận.

4.2 Sigmoid function.

- Sử dụng hàm sigmoid để ánh xạ dự đoán theo xác suất. Đây là công thức của hàm sigmoid:

$$S(z) = \frac{1}{1 + e^{-z}}$$

S(z): là đầu ra giá trị của nó dao động từ 0 đến 1 (ước tính xác suất).
z: là đầu vào tham số thuật toán dự đoán.
e: là số logarit tự nhiên.

- Ranh giới quyết định: Hàm dự đoán hiện tại trả về điểm xác suất trong khoảng từ 0 đến 1, vậy chúng ta cần một ngưỡng để ánh xạ nó thành giá trị true/false

  Nếu P > 0.5, class = 1 có nghĩa là khách hàng chấp nhận sử dụng phiếu giảm giá
  Ngược lại P < 0.5, class = 0 có nghĩa là khách hàng không chấp nhận sử dụng phiếu giảm giá.

- Hàm toán học:

$$P(class = 1) = \frac{1}{1 + e^{-z}}$$

Trong đó: z = W(0) + W(1)*destination + W(2)*passanger +. . . + W(24)*direction_opp
⇒ Nếu hàm dự đoán trả về giá trị là 0.2 tức là xác suất dự đoán là 20%. Ranh giới quyết định ban đầu là 0.5 thì ta sẽ kết luận là 0 có nghĩa là khách hàng đó không chấp nhận sử dụng phiếu giảm giá phiếu giảm giá. Còn ngược lại hàm dự đoán trả 0.7 thì ta kết luận là 1 có nghĩa là khách hàng chấp nhận dử dụng phiếu giảm giá

4.3 Cost function

Không nên sử dụng hàm lỗi MSE(L2) giống như hồi quy tuyến tính. Bởi vì hàm dự đoán là phi tuyến tính. Nếu sử dụng MSE model sẽ là một hàm không lồi với nhiều local minimum thì sẽ khó có thể tối ưu được hàm lỗi.
⇒ Vì vậy chúng ta nên thay thế bằng một hàm mất mát là: Cross-Entropy
Hàm này có thể chia thành 2 hàm lỗi riêng biệt với y = 1 và y = 0

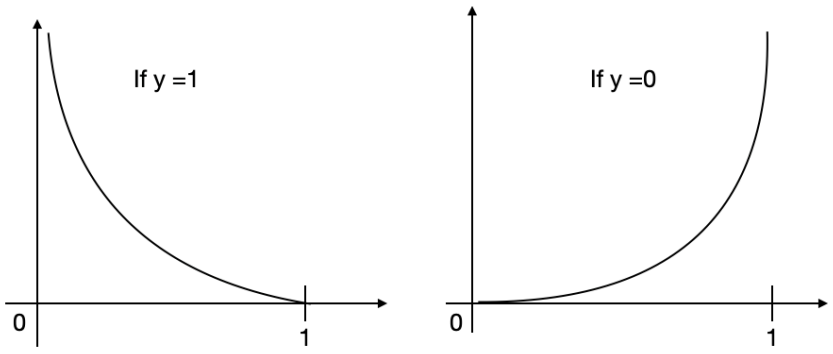
$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y} = 1$$
$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y} = 0$$

J: là hàm lỗi tổng quát và nó củng là trung bình của tất cả các lỗi
Cost(): là hàm chi phí.
h(): là model.
y: là giá trị dự đoán.
Dưới đây là một biến đổi hàm lỗi khi kết hợp y=1 và y=0

$$J(\theta) = \frac{1}{m} \cdot \left(-y^T \log(h) - (1 - y)^T \log(1 - h)\right)$$

4.4 Tối ưu hàm lỗi (Gradient desent)



- Một trong những tính chất của hàm sigmoid là là đạo hàm của nó rất dễ tính

$$S'(z) = S(z)(1 - S(z))$$

- Đạo hàm nó theo weight thì nó sẽ được 1 hàm Cost():

$$C' = x(S(z) - y)$$

  + C': là đạo hàm của hàm chi phí liên quan đến weight.
  + y: các nhãn trong dataset(0 hoặc 1).
  + S(z): là dữ liệu mô hình dự đoán.
  + x là thuộc tính hoặc nhiều thuộc tính của dataset.

CHƯƠNG 3: DỮ LIỆU THỰC NGHIỆM

1 Dataset

1.1 Mô tả dataset

Dataset mô tả về tình huống tuyến đường khách hàng di chuyển và công ty sẽ dựa vào các yếu tố là điểm đến, thời gian, thời tiết, người đi cùng,... và sau đó sẽ phát cho khách hàng phiếu giảm giá và dự đoán xem khách hàng đó có sử dụng mã giảm giá ghé ngang 1 cửa hàng để sử dụng mã giảm giá cho cửa hàng đó hay không.

1.2 Chi tiết thuộc tính

Dataset này nhóm chúng em lấy trên UCI và dataset này có 25 thuộc tính và có 22079 dòng:

Destination	Là điểm đến: No Urgent Place, Home, Work
Passanger	Là người đi cùng: Friend(s), Alone, Kid(s) Partner
Weather	Là thời tiết: Sunny, Rainy, Snowy
Temperature	Là nhiệt độ: 55, 80, 30
Time	Là các giờ cao điểm: 7AM, 10AM, 2PM, 6PM, 10PM
Coupon	Các loại phiếu giảm giá của các cửa hàng: Restaurant<$20, Restaurant(20-50), Coffee House, Bar, Carry out & take away
Expiration	Thời gian sử dụng của phiếu là 2h hoặc 1d
Gender	Giới tính: Female, Male
Age	Độ tuổi: dưới 21, 21,31, 36 46, trên 50
MaritalStatus	Tình trạng hôn nhân: Unmarried partner, Single, Married partner, Divorced(đã ly dị), Widowed(goá chồng/vợ)
has_Children	Đã có con chưa: 1 đã có con, 0 chưa có con
Education	Trình độ học vấn: Some college - no degree(có bằng cao đẳng hoặc không bằng cấp), Bachelors degree(cử nhân), Associates degree(bằng cấp liên kết), High School Graduate, Graduate degree (thạc/tiến sĩ), Some High School
Occupation	Nghề nghiệp: Unemployed, Architecture & Engineering, Student, Education&Training&Library, Healthcare Support,Healthcare Practitioners & Technical, Sales & Related, Management, Arts Design Entertainment Sports & Media, Computer & Mathematical, Life Physical Social Science, Personal Care & Service, Community & Social Services, Office & Administrative Support, Construction & Extraction, Legal, Retired, Installation Maintenance & Repair, Transportation & Material Moving, Business & Financial, Protective Service,Food Preparation & Serving Related, Production Occupations,Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry
Income	Bình quân thu nhập của khách hàng trên năm: Ít hơn $12500, $12500 - $24999, $25000 - $37499, $37500 - $49999, $50000 - $62499, $62500 - $74999, $75000 - $87499, $87500 - $99999, $100000 hoặc hơn
Bar	Số lần khách hàng đến Bar trong tháng: never, less1, 1~3, 4~8, gt8
CoffeHouse	Số lần khách hàng đến Coffee house trong tháng: never, less1, 1~3, 4~8, gt8
CarryAway	Số lần khách hàng mua các loại đồ ăn mang đi trong tháng: never, less1, 1~3, 4~8, gt8.
RestaurantLessThan20	Số lần khách hàng đến nhà hàng có mức giá <20$ bao nhiêu lần trên tháng: never, less1, 1~3, 4~8, gt8.
Restaurant20To50	Số lần khách hàng đến nhà hàng có mức giá 20-50 bao nhiêu lần trên tháng: never, less1, 1~3, 4~8, gt8.
ToCoupon_GEQ5min	Khoảng cách lái xe đến điểm sử dụng mã giảm giá là 5 phút: 1 là đến nơi để sử dụng, 0 là không có
ToCoupon_GEQ15min	Khoảng cách lái xe đến điểm sử dụng mã giảm giá là 15 phút: 1 là đến nơi để sử dụng, 0 là không có
ToCoupon_GEQ25min	Khoảng cách lái xe đến điểm sử dụng mã giảm giá là 25 phút: 1 là đến nơi để sử dụng, 0 là không có
Direction_same	Nhà hàng/ bar có cùng với điểm đến hay không: 1 có, 0 là không
Direction_opp	Nhà hàng/ bar có khác hướng với điểm đến hay không: 1 có, 0 là không
Y	Kết luận khách hàng có sử dụng phiếu giảm giá hay không: 1 có, 0 là không

- Trong đó file Train chiếm 80% là 17,663 và file text chiếm 20% là 4.416. Ta sẽ xóa cột car và xóa các dòng chứa giá trị null.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,random_state = 42)
```

- Dữ liệu bị null

destinatio	passanger	weather	temperat	time	coupon	expiration	gender	age	maritalSta	has_childr	education	occupatio	income	car	Bar	CoffeeHou	CarryAway
No Urgent	Alone	Sunny	55	2PM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	80	10AM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	80	10AM	Carry out	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	80	2PM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	80	2PM	Coffee Ho	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	80	6PM	Restauran	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Friend(s)	Sunny	55	2PM	Carry out	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	80	10AM	Restauran	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	80	10AM	Carry out	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	80	10AM	Bar	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	80	2PM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	55	2PM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Kid(s)	Sunny	55	6PM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Home	Alone	Sunny	55	6PM	Bar	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Home	Alone	Sunny	55	6PM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Home	Alone	Sunny	80	6PM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	55	7AM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	55	7AM	Bar	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	80	7AM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	80	7AM	Carry out	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	55	7AM	Restauran	1d	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
Work	Alone	Sunny	55	7AM	Coffee Ho	2h	Female	21	Unmarriec	1	Some coll	Unemploy	$37500 - $49999		never	never	
No Urgent	Alone	Sunny	55	2PM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 - $74999		never	less1	4~8

- Dữ liệu sau khi xử lý

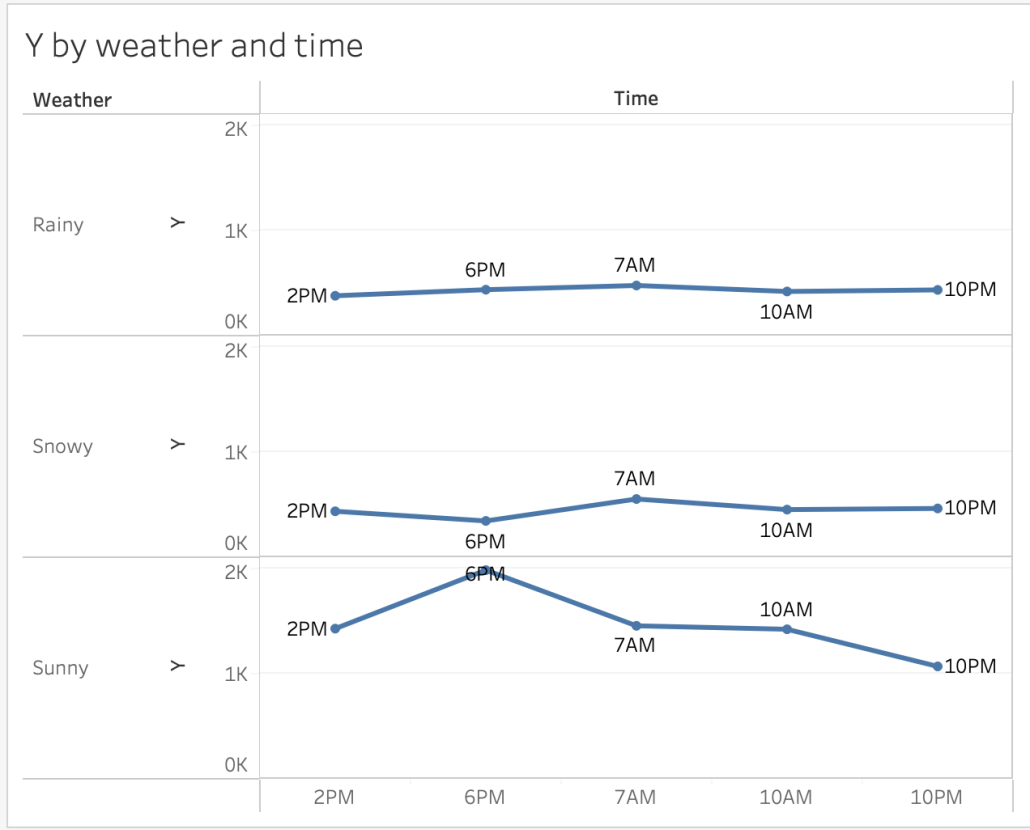
destinatio	passanger	weather	temperati	time	coupon	expiration	gender	age	maritalSta	has_childr	education	occupatio	income	Bar	CoffeeHo	CarryAwa	Restauran	Restauran	toCoupon	toCoupon	toCoupon	direction_	direction_	Y
No Urger	Alone	Sunny	55	2PM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	1
No Urger	Friend(s)	Sunny	80	10AM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
No Urger	Friend(s)	Sunny	80	10AM	Bar	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	1
No Urger	Friend(s)	Sunny	80	10AM	Carry out	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
No Urger	Friend(s)	Sunny	80	2PM	Coffee Ho	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
No Urger	Friend(s)	Sunny	80	2PM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
No Urger	Friend(s)	Sunny	80	2PM	Coffee Ho	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
No Urger	Friend(s)	Sunny	80	2PM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	1
No Urger	Friend(s)	Sunny	80	6PM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
No Urger	Friend(s)	Sunny	80	6PM	Restauran	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	1
No Urger	Friend(s)	Sunny	55	2PM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
No Urger	Friend(s)	Sunny	55	2PM	Carry out	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	1
No Urger	Alone	Sunny	55	10AM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
Home	Alone	Sunny	55	6PM	Bar	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	1	0	1
Home	Alone	Sunny	55	6PM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
Home	Alone	Sunny	80	6PM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	0
Work	Alone	Sunny	55	7AM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	1	0	1	0
Work	Alone	Sunny	55	7AM	Bar	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	1	0	1	1
Work	Alone	Sunny	80	7AM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
Work	Alone	Sunny	80	7AM	Carry out	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	1	0	1
Work	Alone	Sunny	55	7AM	Restauran	1d	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	0	0	0	1	1
Work	Alone	Sunny	55	7AM	Coffee Ho	2h	Male	21	Single	0	Bachelors	Architectu	$62500 -	never	less1	4~8	4~8	less1	1	1	0	0	1	0
No Urger	Alone	Sunny	55	2PM	Restauran	1d	Male	46	Single	0	Some coll	Student	$12500 -	never	4~8	1~3	1~3	never	1	0	0	0	1	1
No Urger	Friend(s)	Sunny	80	10AM	Coffee Ho	2h	Male	46	Single	0	Some coll	Student	$12500 -	never	4~8	1~3	1~3	never	1	0	0	0	1	1
No Urger	Friend(s)	Sunny	80	10AM	Bar	1d	Male	46	Single	0	Some coll	Student	$12500 -	never	4~8	1~3	1~3	never	1	0	0	0	1	0
No Urger	Friend(s)	Sunny	80	10AM	Carry out	2h	Male	46	Single	0	Some coll	Student	$12500 -	never	4~8	1~3	1~3	never	1	1	0	0	1	1

- Hàm dùng để xử lý giá trị null

```
: df.shape
: (12684, 25)
```

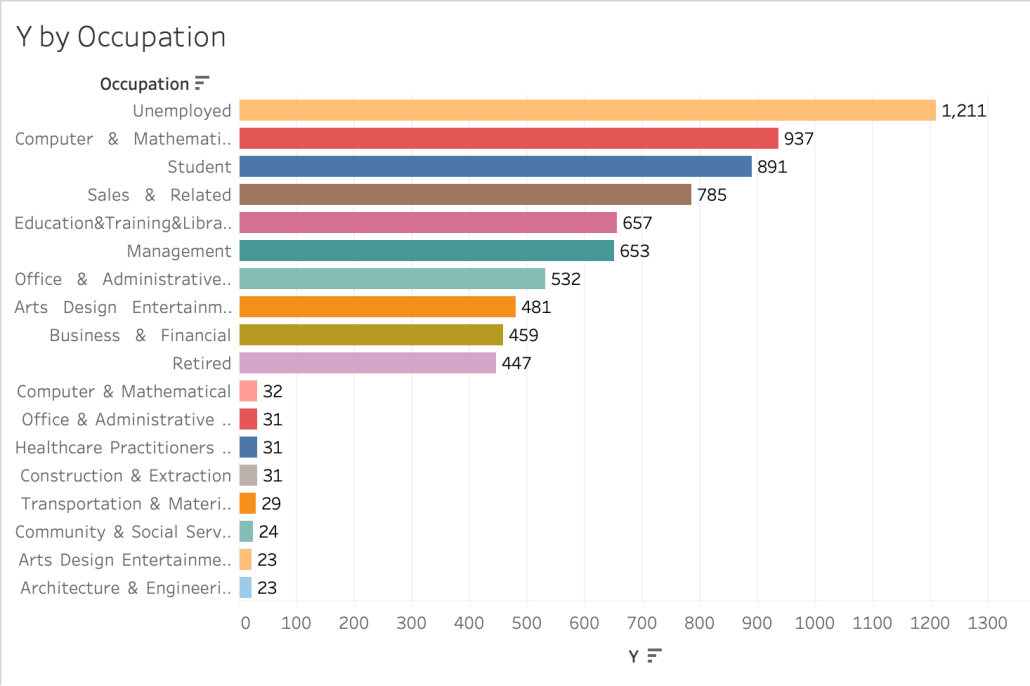
```
: for i in df:
      row=df[df[i].isnull()].index.tolist()
      df = df.drop(labels=row, axis=0)
df.reset_index(drop=True, inplace=True)
df.shape
: (12079, 25)
```

2 Biểu đồ trực quan



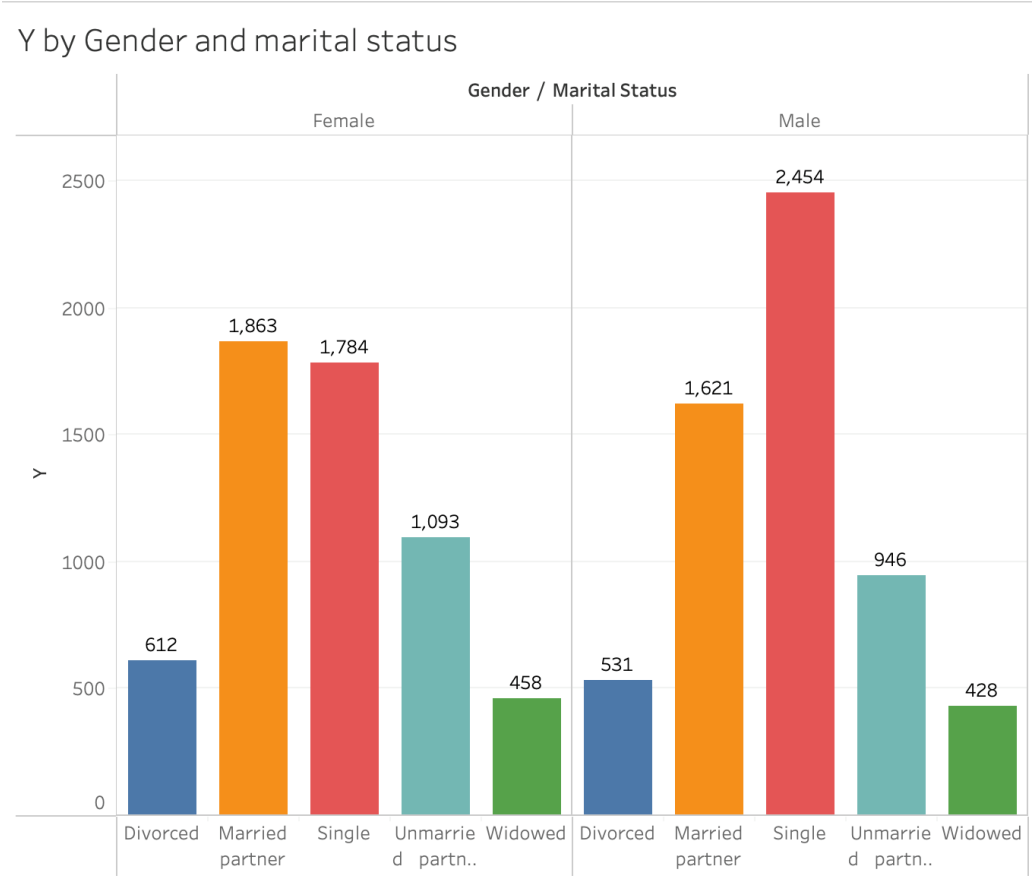
Biểu đồ này thể hiện tổng số phiếu giảm giá khách hàng sử dụng vào các giờ cao điểm theo thời tiết

+ Rainy thì khách hàng sử dụng phiếu giảm giá vào 7AM là nhiều nhất
+ Snowy thì khách hàng sử dụng phiếu giảm giá củng vào 7AM nhiều nhất.
+ Sunny thì khách hàng sử dụng phiếu giảm giá vào 6PM là nhiều nhất.
+ Tỉ lệ khách hàng sử dụng phiếu giảm giá vào thời tiết Sunny của các giờ đều cao hơn của Rainy và Snowy.



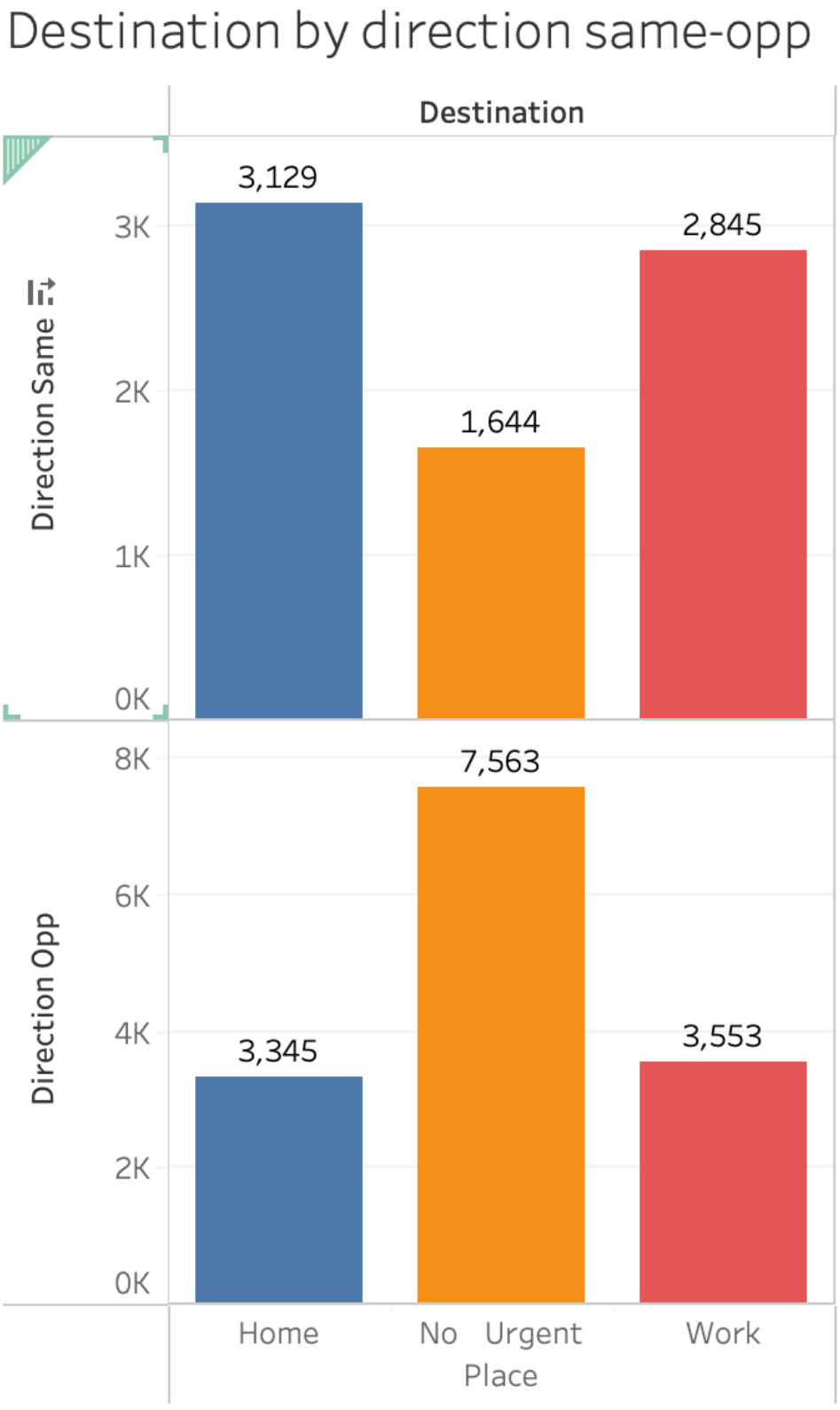
Biểu đồ thể hiện tổng phiếu giảm giá được phát từ công ty và tổng phiếu giảm giá được sử dụng dựa theo nghề nghiệp của họ.

+ Nghề nghiệp Unemployed sử dụng phiếu giảm giá cao nhất là 1211 phiếu.
+ Nghề nghiệp Architecture & Engineering sử dụng thấp nhất là là 196 phiếu.



Biểu đồ thể hiện tình trạng hôn nhân và giới tính của khách hàng nào sử dụng phiếu giảm giá.

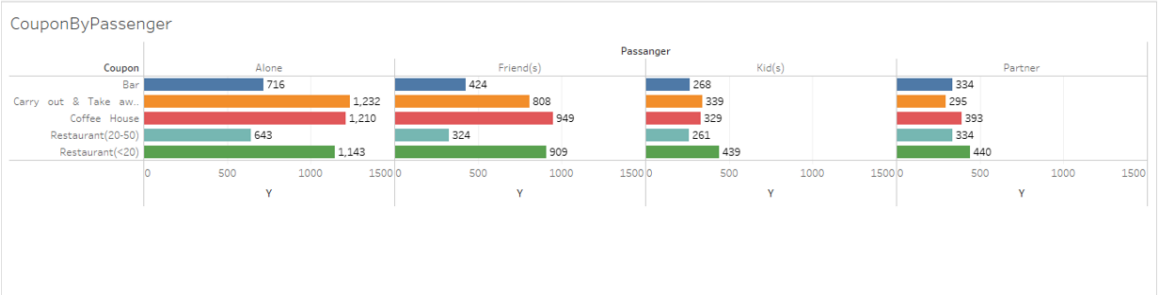
+ Ta thấy tình trạng hôn nhân: đã kết hôn và độc thân của nam và nữ sử dụng phiếu giảm giá được thể hiện cao nhất so với các tình trạng hôn nhân còn lại.



Biểu đồ thể hiện phiếu giảm giá cùng hay khác hướng với điểm đến của khách hàng.

+ Hàng trên thể hiện khách hàng sẽ sử dụng phiếu giảm giá khi nhà hàng/bar cùng hướng với điểm đến là home hoặc work nhiều nhất.

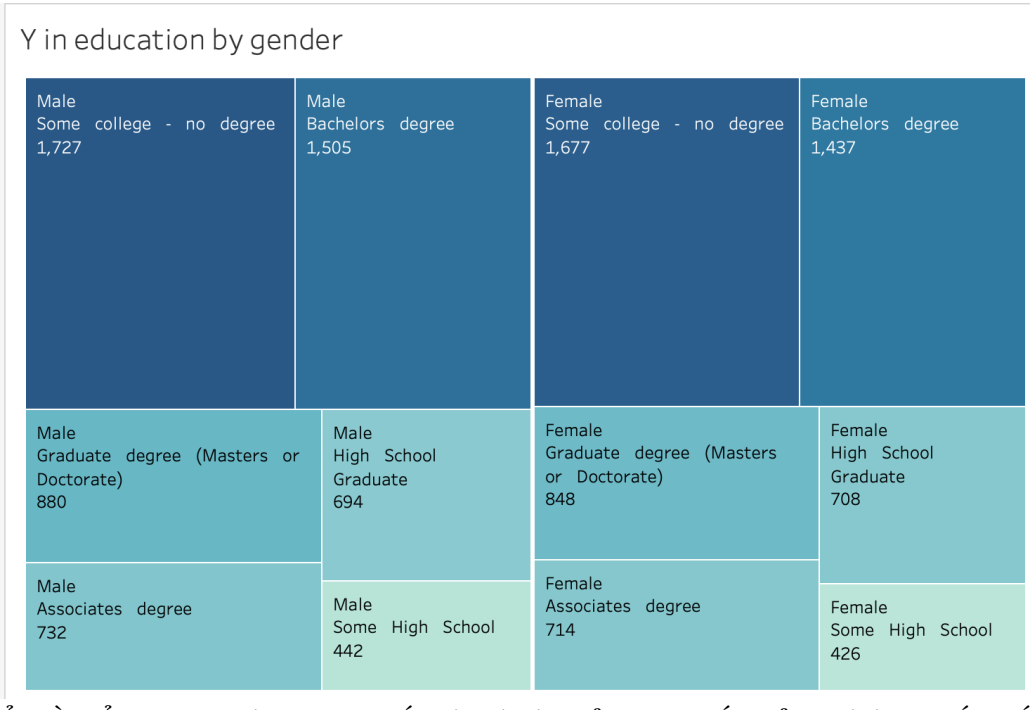
+ Còn hàng dưới thể hiện các khách hàng đi dạo hoặc chưa có điểm đến thì sẽ lựa chọn ngược hướng với điểm đến nhiều nhất.



Bảng này thể hiện các loại phiếu giảm giá được sử dụng bởi khách hàng có người đi cùng.

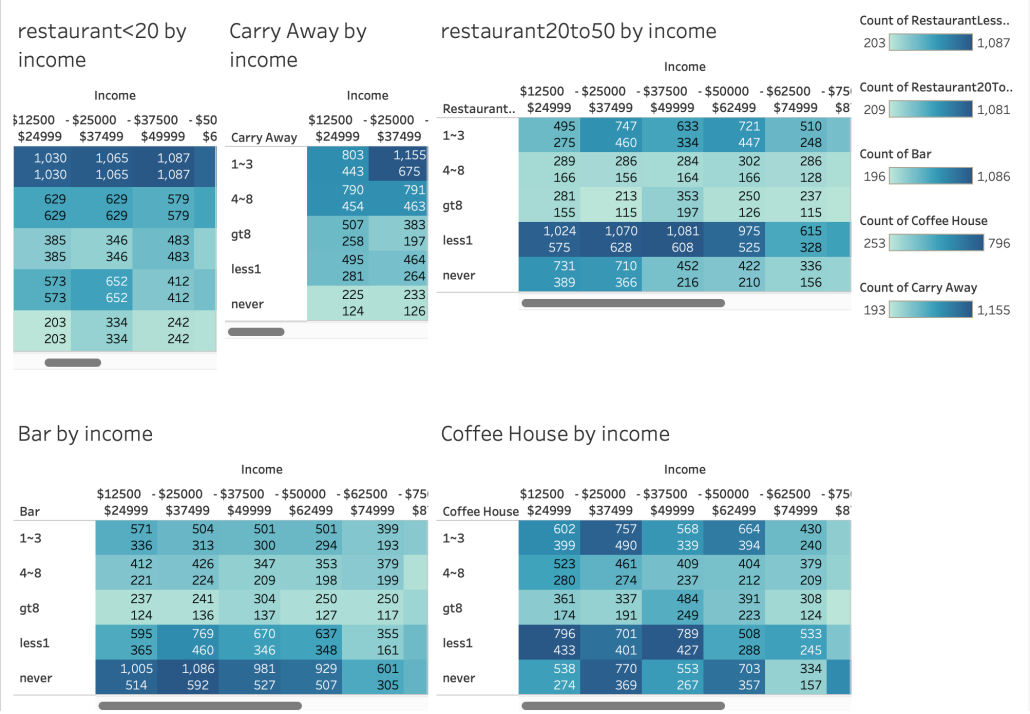
+ Ta thấy đối với người đi 1 mình thì sử dụng phiếu mang đi nhiều nhất.
+ Người đi với bạn thì sử dụng phiếu coupon đến coffee house nhiều nhất.
+ Người có chở trẻ em theo thì sử dụng phiếu coupon đến nhà hàng có mức giá < 20$ nhiều nhất.
+ Người đi cùng đồng nghiệp thì sử dụng phiếu đến nhà hàng có mức giá < 20$ nhiều nhất.

⇒ Các phiếu coupon đa số được khách hàng đi một mình và đi với bạn sử dụng nhiều.



Biểu đồ thể phân loại trình độ học vấn và giới tính sử dụng phiếu giảm giá từ cao đến thấp.

+ Khách hàng có bằng cao đẳng hoặc không bằng cấp với giới tính là nam và nữ sử dụng phiếu giảm giá cao nhất là 1.727 và 1.677.
+ Khách hàng có trình độ trung học và giới tính nam, nữ sử dụng phiếu giảm giá thấp nhất là 442 và 426.



Biểu đồ thể hiện phân khúc khách hàng theo thu nhập của khách hàng.

+ Đây là 5 biểu đồ nhỏ cho từng loại cửa hàng được gộp vào 1 Dashboard để dễ dàng quan sát.

+ Ta thấy rằng rõ rang phân số lượng khách hàng có thu nhập từ 12500-49999 sử dụng phiếu giảm giá đến nhiều nhất.

CHƯƠNG 4: THỰC NGHIỆM

1 Kết quả thực nghiệm

Trong bài nghiên cứu này, chúng tôi chủ yếu tập trung vào bốn tiêu chí đánh giá hiệu suất để so sánh các mô hình: Accuracy, Recall, Precision, F1-score

- Accuracy: Độ chính xác được định nghĩa là tỷ lệ giữa số lượng mẫu được phân loại chính xác bởi bộ phân loại trên tổng số mẫu cho một tập dữ liệu thử nghiệm nhất định. Công thức như sau:

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

- Recall: Recall là một phần của tất cả các trường hợp positive mà các điểm đó chính xác là positive.

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

- Precision: Precision là một phần của tất cả các trường hợp positive mà được phân loại là positive.

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

- F1- Score: Điểm F1, còn được gọi là Điểm F cân bằng, được định nghĩa là mức trung bình cân bằng của Accuracy và Recall. Công thức như sau:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Ta sẽ sử dụng thư viện có sẵn trên python là sklearn để thực nghiệm 4 giải thuật. Nhưng trước hết ta sẽ phải chuẩn hóa dữ liệu để đầu vào của data khớp với đầu vào của 4 giải thuật. Ta sẽ sử dụng module có sẵn của sklearn là Label Encoder cho từng cột thuộc dạng string. Và ta đã có bộ data phù hợp với đầu vào của 4 module

```
df.destination=encoder.fit_transform(df.destination)
df.passanger=encoder.fit_transform(df.passanger)
df.weather=encoder.fit_transform(df.weather)
df.time=encoder.fit_transform(df.time)
df.coupon=encoder.fit_transform(df.coupon)
df.age=encoder.fit_transform(df.age)
df.expiration=encoder.fit_transform(df.expiration)
df.gender=encoder.fit_transform(df.gender)
df.maritalStatus=encoder.fit_transform(df.maritalStatus)
df.education=encoder.fit_transform(df.education)
df.occupation=encoder.fit_transform(df.occupation)
df.income=encoder.fit_transform(df.income)
df.Bar=encoder.fit_transform(df.Bar)
df.CoffeeHouse=encoder.fit_transform(df.CoffeeHouse)
df.CarryAway=encoder.fit_transform(df.CarryAway)
df.RestaurantLessThan20=encoder.fit_transform(df.RestaurantLessThan20)
df.Restaurant20To50=encoder.fit_transform(df.Restaurant20To50)
```

```
df.head()
```

	destination	passanger	weather	temperature	time	coupon	expiration	gender	age	maritalStatus	...	CoffeeHouse	CarryAway	RestaurantLessThan20	Re
0	1	0	2	55	2	4	0	1	0	2	...	3	1	1	
1	1	1	2	80	0	2	1	1	0	2	...	3	1	1	
2	1	1	2	80	0	0	0	1	0	2	...	3	1	1	
3	1	1	2	80	0	1	1	1	0	2	...	3	1	1	
4	1	1	2	80	2	2	0	1	0	2	...	3	1	1	

Tiếp đến ta sẽ áp dụng 4 giải thuật vào dataset và ta sẽ thu được kết quả là ta có được:

_ Áp dụng thuật toán Decision Tree:

Giải thuật Decision Tree

```
from sklearn import tree
dt_clf = tree.DecisionTreeClassifier()
dt_clf.fit(X_train,y_train)
y_pred_dt = dt_clf.predict(X_test)
TP_dt,TN_dt,FP_dt,FN_dt=getCM(y_test,y_pred_dt)
```

ta thu được kết quả

```
df_matrix_dt =confusion_matrix_func(TP_dt,TN_dt,FP_dt,FN_dt)
print(df_matrix_dt)
```

```
                        Actual: Positive  Actual: Negative
0  Predicted as Positive              1464               835
1  Predicted as Negative               933              1184
```

- True Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã thật sự đã dùng (1464)
- False Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã không dùng voucher (835)
- True Negative: Số lượng khách hàng được dự đoán không dùng voucher thì thật sự đã không dùng voucher (1184)
- False Negative: Số lượng khách hàng được dự đoán không dùng voucher thì đã dùng voucher (933).

_ Áp dụng thuật toán Random Forest:

Giải thuật Random Forest

```
from sklearn import ensemble
rf_clf = ensemble.RandomForestClassifier()
rf_clf.fit(X_train,y_train)
y_pred_rf = rf_clf.predict(X_test)
TP_rf,TN_rf,FP_rf,FN_rf=getCM(y_test,y_pred_rf)
```

ta thu được kết quả

```
df_matrix_rf =confusion_matrix_func(TP_rf,TN_rf,FP_rf,FN_rf)
print(df_matrix_rf)
```

		Actual: Positive	Actual: Negative
0	Predicted as Positive	1577	814
1	Predicted as Negative	820	1205

- True Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã thật sự đã dùng (1577)
- False Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã không dùng voucher (814)
- True Negative: Số lượng khách hàng được dự đoán không dùng voucher thì thật sự đã không dùng voucher (1205)
- False Negative: Số lượng khách hàng được dự đoán không dùng voucher thì đã dùng voucher (820).

_ Áp dụng thuật toán Naive Bayes:

Giải thuật Naive Bayes

```
from sklearn import naive_bayes
nb_clf = naive_bayes.GaussianNB()
nb_clf.fit(X_train,y_train)
y_pred_nb = nb_clf.predict(X_test)
TP_nb,TN_nb,FP_nb,FN_nb=getCM(y_test,y_pred_nb)
```

ta thu được kết quả

```
df_matrix_nb =confusion_matrix_func(TP_nb,TN_nb,FP_nb,FN_nb)
print(df_matrix_nb)
```

		Actual: Positive	Actual: Negative
0	Predicted as Positive	1502	1054
1	Predicted as Negative	895	965

- True Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã thật sự đã dùng (1502)
- False Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã không dùng voucher (1054)
- True Negative: Số lượng khách hàng được dự đoán không dùng voucher thì thật sự đã không dùng voucher (965)
- False Negative: Số lượng khách hàng được dự đoán không dùng voucher thì đã dùng voucher (895).

_ Áp dụng thuật toán Logistic Regression:

Giải thuật Logistic Regression

```
from sklearn import linear_model
lr_clf = linear_model.LogisticRegression(solver='lbfgs', max_iter=500)
lr_clf.fit(X_train,y_train)
y_pred_lr = lr_clf.predict(X_test)
TP_lr,TN_lr,FP_lr,FN_lr=getCM(y_test,y_pred_lr)
```

ta thu được kết quả

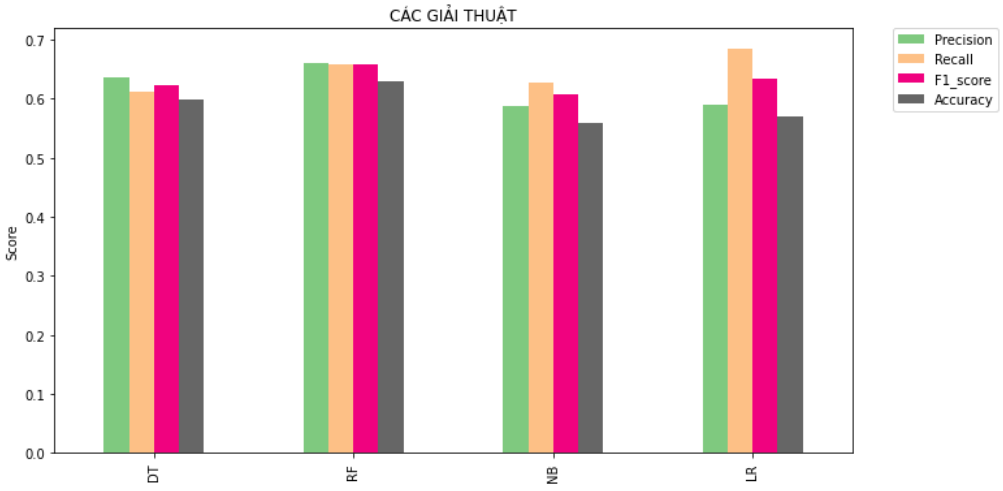
```
df_matrix_lr =confusion_matrix_func(TP_lr,TN_lr,FP_lr,FN_lr)
print(df_matrix_lr)
```

```
                        Actual: Positive  Actual: Negative
0  Predicted as Positive              1642              1145
1  Predicted as Negative               755               874
```

- True Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã thật sự đã dùng (1642)
- False Positive: Số lượng khách hàng được dự đoán dùng voucher thì đã không dùng voucher (1145)
- True Negative: Số lượng khách hàng được dự đoán không dùng voucher thì thật sự đã không dùng voucher (874)
- False Negative: Số lượng khách hàng được dự đoán không dùng voucher thì đã dùng voucher (755).

2 So sánh kết quả

	Name	Precision	Recall	F1_score	Accuracy
0	DT	0.636799	0.610763	0.623509	0.599638
1	RF	0.659557	0.657906	0.658730	0.629982
2	NB	0.587637	0.626617	0.606501	0.558650
3	LR	0.589164	0.685023	0.633488	0.569746



Qua số liệu thu được ta có thể nhận xét rằng: Giải thuật Decision Tree

- Precision: Là tỷ lệ giữa những khách hàng dùng voucher so với tổng số lượng khách hàng được dự đoán là sẽ dùng voucher là 2299 (True Positive + False Positive) và tỷ lệ mô hình đưa ra là 63.7%, tức khoảng 1464 người thật sự dùng voucher.
- Recall: Là tỷ lệ khách hàng dùng voucher mà mô hình dự đoán đúng trong tổng số những khách hàng dùng voucher. Nghĩa là, số lượng khách hàng thật sự dùng voucher là 2397 (True positive + False Negative), recall của mô hình là 61.1% thì kết quả dự đoán là khoảng 1464 khách hàng dùng voucher
- Accuracy: Là độ chính xác của mô hình đào tạo. Mô hình dự đoán của chúng tôi có độ chính xác là 59.9%.

Giải thuật Random Forest

- Precision: Là tỷ lệ giữa những khách hàng dùng voucher so với tổng số lượng khách hàng được dự đoán là sẽ dùng voucher là 2319 (True Positive + False Positive) và tỷ lệ mô hình đưa ra là 66%, tức khoảng 1578 người thật sự dùng voucher.
- Recall: Là tỷ lệ khách hàng dùng voucher mà mô hình dự đoán đúng trong tổng số những khách hàng dùng voucher. Nghĩa là, số lượng khách hàng thật sự dùng voucher là 2397 (True positive + False Negative), recall của mô hình là 65.8% thì kết quả dự đoán là khoảng 1577 khách hàng dùng voucher
- Accuracy: Là độ chính xác của mô hình đào tạo. Mô hình dự đoán của chúng tôi có độ chính xác là 63%.

Giải thuật Naive Bayes

- Precision: Là tỷ lệ giữa những khách hàng dùng voucher so với tổng số lượng khách hàng được dự đoán là sẽ dùng voucher là 2556 (True Positive + False Positive) và tỷ lệ mô hình đưa ra là 58.8%, tức khoảng 1502 người thật sự dùng voucher.
- Recall: Là tỷ lệ khách hàng dùng voucher mà mô hình dự đoán đúng trong tổng số những khách hàng dùng voucher. Nghĩa là, số lượng khách hàng thật sự dùng voucher là 2397 (True positive + False Negative), recall của mô hình là 62.7% thì kết quả dự đoán là khoảng 1502 khách hàng dùng voucher
- Accuracy: Là độ chính xác của mô hình đào tạo. Mô hình dự đoán của chúng tôi có độ chính xác là 55.9%.

Giải thuật Logitic Regression

- Precision: Là tỷ lệ giữa những khách hàng dùng voucher so với tổng số lượng khách hàng được dự đoán là sẽ dùng voucher là 2787 (True Positive + False Positive) và tỷ lệ mô hình đưa ra là 58.9%, tức khoảng 1641 người thật sự dùng voucher.
- Recall: Là tỷ lệ khách hàng dùng voucher mà mô hình dự đoán đúng trong tổng số những khách hàng dùng voucher. Nghĩa là, số lượng khách hàng thật sự dùng voucher là 2397 (True positive + False Negative), recall của mô hình là 68.5% thì kết quả dự đoán là khoảng 1641 khách hàng dùng voucher
- Accuracy: Là độ chính xác của mô hình đào tạo. Mô hình dự đoán của chúng tôi có độ chính xác là 57%.

3 Kết luận

Trong 4 giải thuật này thì Random Forest cho kết quả dự đoán cao nhất với độ chính xác tổng quát cao nhất là 63%

CHƯƠNG 5: KẾT LUẬN

1 Những đóng góp chính của báo cáo và thành quả đạt được

1. Những đóng góp chính
    - Tổng quan được bài toán
    - Trình bày và giới thiệu 4 thuật toán là Decision Tree, Random Forest, Naive Bayes, Logistic Regression
    - Giải thích các thuộc tính và trình bày dữ liệu một các trực quan nhất trên phần mềm Tableau
    - Áp dụng và đánh giá 4 giải thuật đã nêu trên
2. Thành quả đạt được
    - Giới thiệu rõ ràng và thiết thực vấn đề vì sao cần phải Dự đoán việc sử dụng voucher trên ứng dụng thanh toán.
    - Trình bày rõ ràng và đầy đủ, áp dụng được 4 giải thuật để làm ví dụ minh họa lên bộ dataset
    - Trình bày các thuộc tính có trong bộ dataset và đã trực quan hóa dữ liệu trên tableau và nêu lên nhận xét được những gì biểu đồ thể hiện ra.
    - Áp dụng thành công 4 giải thuật thông qua module của thư viện sklearn. Đánh giá và giải thích các số liệu như True Positive, False Positive, True Negative, False Negative và các điểm như Accuracy, Precision, Recall, F1-score.

2 Định hướng nghiên cứu trong tương lai

Trong tương lai sẽ mở rộng nghiên cứu hơn đến nhiều đối tượng để làm tăng kích thước của bộ dữ liệu, thêm một số thuộc tính cần thiết để có thể đánh giá chính xác hơn. Áp dụng nhiều giải thuật để đưa ra kết luận chính xác. Nghiên cứu thêm về các trực quan hóa trên Tableau để đưa trình bày nhiều biểu đồ hơn và nhận xét đầy đủ hơn.