**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC BÀ RỊA – VŨNG TÀU**

**KHOA CÔNG NGHỆ KỸ THUẬT - NÔNG NGHIỆP CÔNG NGHỆ CAO**

🙢🕮🙠



**BÁO CÁO MÔN HỌC**

**Trí tuệ nhân tạo**

Sinh viên thực hiện: Vi Hoàng Gia

MSSV: 18033885

Lớp: DH19CT

Ngành: Công nghệ thông tin

**Vũng Tàu, Tháng 12 2022**

**MỤC LỤC**

[Nhận xét của giảng viên 4](#_Toc123146453)

[CHƯƠNG I: Giới thiệu 5](#_Toc123146454)

[I. Máy học và những ứng dụng của máy học 5](#_Toc123146455)

[II. Ứng dụng của Machine Learning trong thực tiễn 6](#_Toc123146456)

[CHƯƠNG II: Thuật toán phân loại và phân tích dữ liệu 8](#_Toc123146457)

[I. Thuật toán phân loại 8](#_Toc123146458)

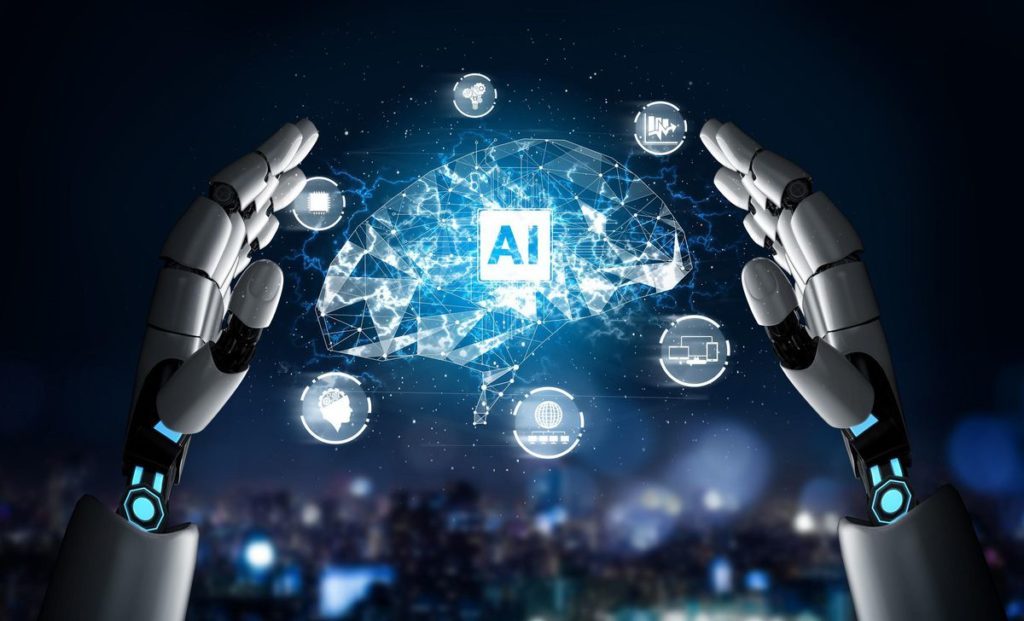
[II. Phân tích dữ liệu 8](#_Toc123146459)

# Nhận xét của giảng viên

# CHƯƠNG I: Giới thiệu

## I. Máy học và những ứng dụng của máy học

Thuật ngữ học máy có lẽ không còn quá xa lạ với con người ngày nay bởi chúng đã được sử dụng trên phạm vi toàn thế giới. Hiện tại, Machine Learning đã đạt được rất nhiều thành tựu to lớn và trở thành hướng nghiên cứu chính trong AI (trí tuệ nhân tạo). Vậy học máy là gì? Đừng bỏ lỡ bài viết sau, chúng mình sẽ giúp bạn tìm hiểu những thông tin liên quan đến Machine Learning.



**Học máy là gì?**

Học máy hay máy học có tên tiếng Anh đầy đủ là Machine Learning, viết tắt là ML. Thuật toán Machine Learning là các chương trình máy tính có khả năng học hỏi và hoàn thành các nhiệm vụ, đồng thời là cách để cải thiện hiệu suất theo thời gian vô cùng hiệu quả.

Ngoài ra, Machine Learning còn được biết là còn là công nghệ phát triển từ lĩnh vực trí tuệ nhân tạo tiên tiến nhất hiện nay. Để đảm bảo không có sai lệch và không xuất hiện dữ liệu giả, Machine Learning vẫn cần quá trình tìm hiểu và lựa chọn kỹ thuật phân tích dữ liệu từ con người.

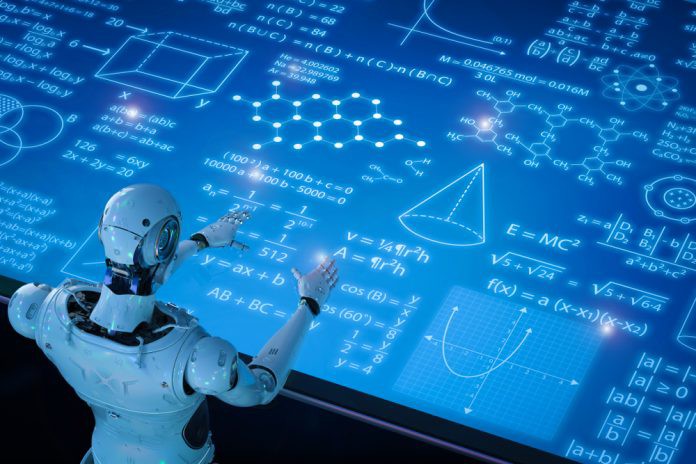
Học máy ngày càng mang tính phổ biến trên toàn thế giới. Sự tăng trưởng vượt bậc của dữ liệu lớn (Big Data) và các thuật toán Machine Learning đã cải thiện độ chính xác của những mô hình và dự đoán tương lai.

**Có bao nhiêu loại học máy ?**

Thuật toán học máy được chia thành 2 loại chính bao gồm: học có giám sát và học không giám sát.

* Học có giám sát (Supervised Learning)

Học có giám sát là phương pháp sử dụng những dữ liệu được gán nhãn sẵn để suy luận ra quan hệ giữa đầu vào và đầu ra. Sau khi tìm hiểu cách tốt nhất để mô hình hóa các mối quan hệ cho dữ liệu được gán nhãn, thuật toán huấn luyện sẽ được sử dụng cho các bộ dữ liệu mới. Ứng dụng của học có giám sát chính là giúp xác định tín hiệu tốt nhất để dự báo xu hướng, lợi nhuận trong tương lai trong lĩnh vực cổ phiếu, chứng khoán.

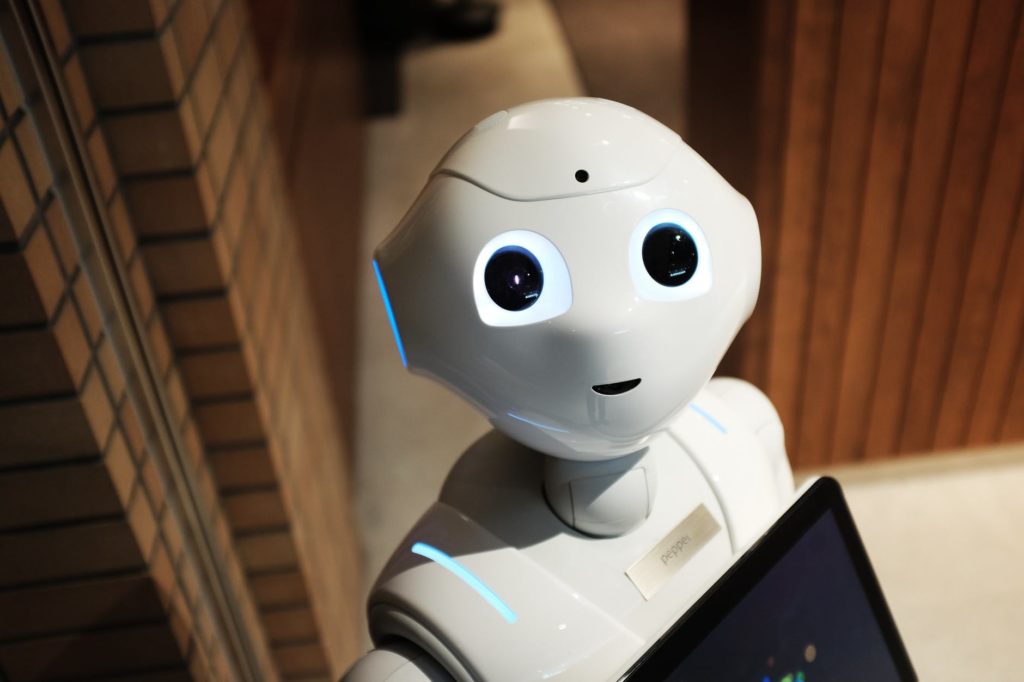


* Học không có giám sát (Unsupervised Learning)

Học không giám sát sử dụng những dữ liệu chưa được gán nhãn sẵn để suy luận và tìm cách để mô tả dữ liệu cùng cấu trúc của chúng. Ứng dụng của học không giám sát đó là hỗ trợ phân loại thành các nhóm có đặc điểm tương đồng.

## II. Ứng dụng của Machine Learning trong thực tiễn

**Học máy** ngày càng được ứng dụng rộng rãi, đa lĩnh vực trong thực tiễn cuộc sống con người ngày nay. Các bạn có thể dàng nhận thấy Machine Learning đang được sử dụng với mục đích phân tích dữ liệu lớn để có thể đưa ra những dự đoán xu hướng trong tương lai. Ví dụ như dự đoán kết quả bầu cử chính trị, dự đoán biến động của thị trường chứng khoán …



Machine Learning cũng được áp dụng trong công nghệ nhận diện hình ảnh nhằm cung cấp dữ liệu cho các mô hình định giá hoặc mô hình kinh tế. Bằng khả năng phân tích dữ liệu từ hệ thống vệ tinh, thuật toán học máy sẽ tiến hành xử lý và cung cấp tới bạn những thông tin về số lượng khách hàng tại bãi đậu xe, các hoạt động vận chuyển, sản lượng sản xuất …

Machine Learning ngày càng thể hiện được rõ vai trò của mình trong sự phát triển của công nghệ trong thời đại mới. Am hiểu về học máy sẽ giúp bạn có thể mở ra những cơ hội phát triển cùng sự thăng tiến trong tương lai. Do đó, bạn hãy xây dựng cho mình một lộ trình nghiên cứu học máy từ cơ bản đến nâng cao để nâng cao trình độ chuyên môn của mình tốt hơn nhé.

# CHƯƠNG II: Thuật toán phân loại và phân tích dữ liệu

## I. Thuật toán phân loại

* Logistic Regression

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

* K-Nearest Neighbor

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based hay Memory-based learning.

* Decision Tree

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

## II. Dữ liệu

1. Bài toán kinh doanh

* Dự đoán liệu khách hàng có chấp thuận phiếu giảm giá hay không là một bài toán khó và chúng ta không thể chỉ giới thiệu nó cho tất cả mọi người vì các chi phí liên quan. Vì vậy, trong bài toán này, chúng ta sẽ dự đoán xem khách hàng sẽ chấp thuận hay từ chối phiếu giảm giá được cung cấp dựa trên hồ sợ và lịch sử khách hàng.

1. Yêu cầu

* Mục tiêu của bài toán là dự đoán xem khách hàng sẽ chấp nhận hay từ chối phiếu giảm giá một địa điểm cụ thể dựa trên các thuộc tính. Nếu khách hàng chấp nhận phiếu giảm giá sẽ dán nhãn Y=1 và nếu khách hàng từ chối phiếu giảm giá sẽ được đánh giá Y=0.

1. Dữ liệu

* Dữ liệu này được thu thập thông qua một cuộc khảo sát trên Amazon Mechanical Turk được xuất bản trên trang web UCI vào năm 2017. Cuộc khảo sát mô tả các tình huống lái xe khác nhau bao gồm điểm đến người dùng, thời gian, thời tiết, hành khách, thuộc tính phiếu giảm giá,… Sau đó hỏi xem người dùng có chấp thuận phiếu giảm giá không.

Tệp dữ liệu: in-vehicle-coupon-recommendation.csv

Bộ dữ liệu bao gồm hơn 12684 dòng dữ liệu và 26 đặc trưng

Giải thích về các thuộc tính/ đặc trưng của bộ dữ liệu:

**Thuộc tính người dùng:**

- Gender: Female, Male

- Age:21, 46, 26, 31, 41, 50plus, 36, below21

- Marital Status: Unmarried partner, Single, Married partner, Divorced, Widowed

- has\_Children: 1: has children, 0: No children

- Education: Some college — no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School

- Occupation: nghề nghiệp của người dùng (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: thu nhập của người dùng (Less than $12500,$12500 — $24999,$25000 — $37499,$37500 — $49999,$50000 — $62499,$62500 — $74999,$75000 — $87499,$87500 — $99999,$100000 or More)

- Car : Mô tả phương tiện do người dùng điều kiển (Scooter and motorcycle, crossover, Mazda5) (99% of values are missing in this feature)

- Bar: người dùng đến bar bao nhiêu lẩn mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)

- CoffeeHouse: người dùng đến quán cà phê bao nhiêu lần mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)

- CarryAway: người dùng nhận đồ ăn mang về bao nhiêu lần mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)

- RestaurantLessThan20: bao nhiêu lần người dùng đến nhà hàng với chi phí trung bình cho mỗi người dưới 20$ mỗi tháng? (never, less1, 1~3, 4~8, gt8, nan)

- Restaurant20To50: người dùng đến nhà hàng bao nhiêu lần với chi phí trung bình cho mỗi ngườ là 20 – 50$ mỗi tháng? (never, less1, 1~3, 4~8, gt8, nan)

bình cho mỗi ngườ là 20 – 50$ mỗi tháng? (never, less1, 1~3, 4~8, gt8, nan)

**Thuộc tính ngữ cảnh:**

- Destination: điểm đến của người dùng (No Urgent Place, Home, Work)

- Passenger: khách hàng trên xe là ai (Alone, Friend(s), Kid(s), Partner)

- Weather: thời tiết khi người dùng đang lái xe (Sunny, Rainy, Snowy)

- Temperature: nhiệt độ tính bằng độ F khi người dùng lái xe (55, 80, 30)

- Time: thời gian khi người dùng lái xe (2PM, 10AM, 6PM, 7AM, 10PM)

- toCoupon\_GEQ5min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 5 phút (0,1)

- toCoupon\_GEQ15min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 15 phút (0,1)

- toCoupon\_GEQ25min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 25 phút (0,1)

- direction\_same: liệu nhà hàng/ quán cà phê/ quán bar có cùng hướng với điểm đến hiện tại của người dùng hay không (0,1)

- direction\_opp: liệu nhà hàng/ quán cà phê/ quán bar có ở hướng ngược lại với điểm đến hiện tại của người dùng hay không (0,1)

**Thuộc tính phiếu giảm giá**

* Coupon: loại phiếu giảm giá do công ty cung cấp (Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50))

Ở đây, Restaurant(<$20) là mức trả trung bình cho mỗi người dùng dưới $20, nhà hàng không quá đắt. Restaurant($20-$50) có nghĩa là mức trả trung bình cho mỗi người dùng là từ $20 đến $50, nhà hàng hơi đắt tiền

* Expiration: Phiếu giảm giá hết hạn sau 1 ngày hoặc sau 2h (1d, 2h)

**Thuộc tính đích:**

* Y: phiếu giảm giá được chấp thuận hay từ chối, 1: được chấp thuận , 0: bị từ chối

## III. Phân tích dữ liệu

1. Thư viện

Text

Description automatically generated

* Thư viện pandas để đọc, ghi, xử lý file dữ liệu
* Thư viện Matplotlib dùng để vẽ đồ thị

1. Tải dữ liệu

Để tải dữ liệu, chúng ta chỉ cần tệp in-vehicle-coupon-recommendation.csv và tải tệp này bằng khung dữ liệu pandas.

Graphical user interface, application

Description automatically generated Table

Description automatically generated

Bộ dữ liệu này có 12684 dòng dữ liệu và 26 đặc trưng

1. Phân bổ coupon

Graphical user interface, text, application, Word

Description automatically generated

Tập dữ liệu này được cân bằng một phần với nhãn loại chấp nhận là khoảng 57% và loại từ chối là khoảng 43%

1. Loại bỏ trùng lặp

Graphical user interface, text, application

Description automatically generated

1. Kiểm tra các dữ liệu còn thiếu trong tệp dữ liệu

Graphical user interface, application

Description automatically generated with medium confidence

Chúng ta có thể thấy rằng có tổng 6 thuộc tính giá trị bị thiếu, từ tính năng ‘car’ có 99% giá trị bị thiếu nên chúng ta cần bỏ đặc trưng này vì ngay cả sau khi dự đoán các giá trị bị thiếu, đặc trưng này ít quan trọng hơn và nó ít có khả năng dự đoán hơn.

Các đặc trưng khác như ‘Bar’, ‘CoffeeHouse’, ‘CarryAway’, ‘RestaurantLessThan20’, ‘Restaurant20To50’ có khoảng 1% giá trị bị thiếu, vì vậy chúng tôi cần điền các giá trị bị thiếu này bằng 1 số giá trị khác, vì vậy chúng tôi sẽ sử dụng mode imputation.

1. Điền các giá trị bị thiếu bằng cách sử dụng mode imputation

Mode imputation thay thế các giá trị còn thiếu của một đối tượng phân loại bằng giá trị phổ biến nhất của đối tượng đó.

Text

Description automatically generated

1. Mối tương quan giữa các đặc trưng

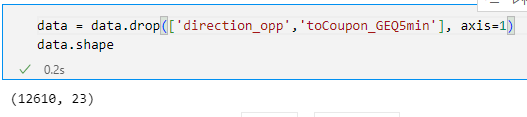
Table

Description automatically generated

Từ ma trận tương quan, chúng ta có thể thấy rằng đặc trưng ‘direction\_same’ tương quan hoàn hảo với ‘direction\_opp’, cả hai đều có cùng giá trị tương quan với thuộc tính đích. Vì vậy không cần 2 tính năng tương quan hoàn hảo để chúng tôi có thể loại bỏ một trong số chúng.

Đặc trưng ‘toCoupon\_GEQ5min’ không tương quan với thuộc tính đích vì đặc trưng này có cùng giá trị ‘1’ cho tất cả dòng dữ liệu, có nghĩa tất cả các điểm ăn uống đều các người dùng ít nhất 5 phút.

Vì vậy, cần bỏ cả hai đặc tính ‘direction\_opp’ và ‘toCoupon\_GEQ5min’ vì đặc tính này ít quan trọng hơn và sẽ không giúp ích nhiều cho việc dự đoán.



1. Phân tích đơn biến

Vì tất cả các thuộc tính đều có tính phân loại, nên chúng ta sẽ xem xét các phân phối kín đáo của các biến khác nhau với biến mục tiêu.

Text, letter

Description automatically generated

1. Coupon:

Logo, company name

Description automatically generated Chart, bar chart

Description automatically generated Table

Description automatically generated

* Chúng ta có thể thấy rằng các phiếu giảm giá được cung cấp nhiều nhất là dành cho Coffee House. Ngoài ra, phiếu giảm giá được nhiều người chấp nhận là phiếu giảm giá ‘Carry out & Take away’ với 73% và ‘Restaurant(<20)’ là 70%. Phiếu giảm giá có thỉ lệ chấp thuận thấp nhất là ‘Bar’

1. Age:

Chart, bar chart

Description automatically generatedTable

Description automatically generated

* Chúng ta có thể thấy rằng hầu hết người dùng trong dữ liệu này có độ tuổi từ 21 đến 31 tuổi. Ngoài ra, người dùng có độ tuổi dưới 21 có tỷ lệ chấp nhận phiếu giảm giá cao nhất. Ngoài ra, người dùng trên 50 tuổi có tỷ lệ từ chối phiếu giảm giá cao nhất

1. InCome:

Chart, bar chart

Description automatically generated

Table

Description automatically generated

* Hầu hết người dùng trong dữ liệu này có thu nhập từ 12500 đô đến 49999 đô. Người dùng có thu nhập cao chấp nhận nhiều phiếu giảm giá hơn những người khác. Ngoài ra người có thu nhập tầm trung hầu hết từ chối phiếu giảm giá.

1. Direction\_same

Chart, bar chart, treemap chart

Description automatically generatedTable

Description automatically generated

* Đặc tính ‘direction\_same’ có hai giá trị là 0 và 1. Cả hai giá trị này có tỷ lệ chấp nhận và từ chối gần như tương tự nhau. Vì vậy đặc tính này không hữu ích trong dự đoán. Vì vậy, bỏ đặc tính này.

1. Phân tích hai biến số
2. Người dùng đi đến ‘Work’, ‘Home’, ‘No Urgent Place’ lúc nào , với ai ?

Text

Description automatically generatedChart, bar chart

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated

* Người dùng đi cùng bạn bè ‘Friend(s)’ chỉ có điểm đến không khẩn cấp ‘No Urgent Place’.
* Vào lúc 7AM, chỉ người dùng đi làm ‘Work’ mới ra ngoài một mình.
* Vào lúc 2PM, người dùng chỉ đến nơi không khẩn cấp ‘No Urgent Place’
* ‘Home’, người dùng chủ yếu chỉ đi 1 mình vào lúc 6PM và 10PM
* ‘No Urgent Place’, người dùng chủ yếu đi một mình ‘Alone’ lúc 2PM và với bạn ‘Friend(s)’ lúc 6PM
* Người dùng đi cùng bạn bè ‘Friend(s)’ lúc 6PM có tỉ lệ chấp nhận cao nhất
* Người dùng đi một mình ‘Alone’ lúc 2PM cũng có tỉ lệ chấp nhận cao
* Người dùng đi cùng trẻ em lúc 10PM hầu hết từ chối phiếu giảm giá

1. Tỷ lệ chấp nhận và từ chối phiếu giảm giá cao vào thời điểm nào

Text

Description automatically generated Chart, bar chart

Description automatically generated Table

Description automatically generated

- Vào lúc 7AM và 2PM, người dùng sẽ không nhận phiếu giảm giá bar có thời hạn 2h

- Vào lúc 2PM và 6PM, người dùng chủ yếu chấp nhận phiếu giảm giá ‘Restaurant(<20)’ và ‘Carry out & take away’ với tỉ lệ 80%

- Vào lúc 2PM, người dùng chủ yếu chấp nhận phiếu giảm giá ‘Carry out & take away’ với tỉ lệ 86%

- Từ 7AM đến 6PM. Phiếu giảm giá Coffee Hosue là nhiều nhất

- 10PM chỉ có Carry out & take away(1d) và Restaurant(<20)(2h) là cao nhất.

- Vào 10PM, người dùng không chấp nhận phiếu giảm giá ‘Restaurant(20-50)’

- Vào 7AM, 10AM, 2PM người dùng từ chối phiếu giảm giá cho Bar

1. Thiết kế đặc trưng
2. To\_Coupon

‘toCoupon\_GEQ15min’ và ‘toCoupon\_GEQ25min’ cho biết khoảng cách lái xe đến nhà hàng/quán cà phê/quán bar, vì vậy chúng tôi có thể kết hợp chúng và tạo một đặc trưng mới gọi là 'to\_coupon'.

Các đặc trưng 'toCoupon\_GEQ15min' và 'toCoupon\_GEQ25min' có hai giá trị phân loại 0 & 1 và 'to\_coupon' có ba giá trị phân loại 0,1 & 2.

0: khoảng cách lái xe nhỏ hơn hoặc bằng 15 phút

1: khoảng cách lái xe lớn hơn 15 phút và nhỏ hơn hoặc bằng 25 phút

2: khoảng cách lái xe lớn hơn 25 phút



Table

Description automatically generated

Chart, bar chart

Description automatically generatedTable

Description automatically generated

Chúng ta có thể thấy rằng hầu hết các phiếu giảm giá được cung cấp đều ở gần đó trong vòng 25 phút. Ngoài ra, người dùng thường chấp nhận các phiếu giảm giá có khoảng cách dưới 25 phút.

1. Coupon\_freq

'coupon\_freq' là sự kết hợp của năm tính năng, RestaurantLessThan20, CoffeeHouse, CarryAway, Bar và Restaurant20To50. Năm tính năng này cho biết tần suất người dùng đến RestaurantLessThan20, CoffeeHouse, CarryAway, Bar và Restaurant20To50.

Mỗi người dùng được yêu cầu một phiếu giảm giá, vì vậy chúng tôi chỉ cần tần suất của một loại phiếu giảm giá đó, không cần các chi tiết phiếu giảm giá khác. vì vậy chúng tôi tạo một tính năng mới cho biết tần suất người dùng sử dụng phiếu giảm giá đã hỏi.

Text

Description automatically generated

Chart, bar chart

Description automatically generated

* Hầu hết người dùng trong dữ liệu này đã truy cập các dịch vụ Nhà hàng(<$20) và Mang đi nhiều lần và họ cũng có tỷ lệ chấp nhận phiếu giảm giá cao nhất.
* Người dùng đã truy cập Coffee House 1 đến 3 lần và 4 đến 8 lần cũng có tỷ lệ chấp nhận phiếu giảm giá cao hơn.
* Hầu hết người dùng chỉ ghé thăm Nhà hàng đắt tiền một lần và những người dùng đã ghé thăm Nhà hàng đắt tiền hơn 4 lần có tỷ lệ chấp nhận phiếu giảm giá cao nhất cho Nhà hàng đắt tiền.
* Hầu hết người dùng trong dữ liệu này chưa bao giờ ghé thăm một Quán bar và hầu hết trong số họ từ chối phiếu giảm giá của quán bar và những người dùng đã ghé thăm Quán bar 1 đến 3 lần hoặc 4 đến 8 lần có tỷ lệ chấp nhận phiếu giảm giá cao hơn.

***Điều đó có nghĩa là coupon\_freq có liên quan nhiều đến biến mục tiêu.***

1. occupation\_class

Đặc điểm occupation\_class có 25 không có giá trị riêng biệt, điều này tạo ra rất ít trong ma trận dữ liệu sau khi mã hóa nên chúng tôi có thể cố gắng chuyển đổi nó thành một đặc điểm mới bằng cách sử dụng kỹ thuật đặc trưng.

Chúng tôi sẽ thử các biến thể khác nhau của việc phân loại các biến số nghề nghiệp, một là chia 25 nghề nghiệp thành năm loại từ kiến ​​thức lĩnh vực hoặc nghề nghiệp của họ nhưng cách đó không hiệu quả lắm. vì vậy chúng tôi sẽ chia chúng thành năm loại cao, medium\_high, medium, medium\_low và low theo thứ tự tỷ lệ chấp nhận phiếu giảm giá và tạo một tính năng mới có tên là class\_class.

Text

Description automatically generated

Chart, bar chart

Description automatically generated

1. Mã hóa dữ liệu

Hầu hết các thuật toán học máy không thể hoạt động trực tiếp trên dữ liệu phân loại. Chúng yêu cầu tất cả các biến đầu vào và biến đầu ra phải là số. Hầu hết các hiệu suất của thuật toán khác nhau dựa trên phương pháp mã hóa nào được sử dụng. Có nhiều cách để mã hóa dữ liệu phân loại.

1. Mã hóa tần số (Frequency Encoding)

Giá trị thay thế nó là tần suất giá trị đó xuất hiện trong cột

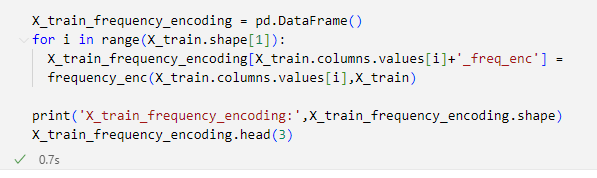
Trong mã hóa này, chúng tôi sẽ chuyển đổi các đặc trưng phân loại thành giá trị tần số của các danh mục của chúng.

Dưới đây là các bước để tính giá trị tần số của tính năng:

1. Nhóm theo biến phân loại và thu được số lượng của từng loại

2. Để chuẩn hóa chúng chia với tổng số của một biến phân loại

3. Bản đồ giá trị tần số với các tính năng đào tạo



Table

Description automatically generated Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence Table

Description automatically generated

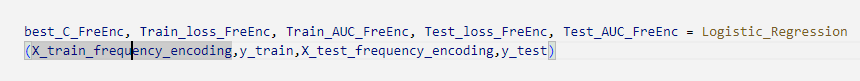
1. Models

Sau khi thực hiện phân tích dữ liệu thăm dò và làm sạch dữ liệu, chúng tôi cũng đã thực hiện một số kỹ thuật đặc trưng và thêm đặc trưng mới 'to\_coupon', 'coupon\_freq' và 'occupation\_class' từ các đặc trưng hiện có. Chúng tôi cũng đã chuyển đổi các tính năng phân loại thành các tính năng số bằng cách sử dụng năm phương pháp mã hóa khác nhau và chuẩn bị năm ma trận dữ liệu kiểm tra và huấn luyện khác nhau. Hiện chúng tôi đã sẵn sàng áp dụng thuật toán học máy cho ma trận dữ liệu đã chuẩn bị của mình.

Có nhiều thuật toán học máy để thử nghiệm khi bạn đang giải quyết vấn đề phân loại. Bạn không bao giờ biết trước mô hình phân loại nào sẽ phù hợp nhất với mình. Vì vậy, chúng tôi sẽ thử nghiệm với một vài mô hình phân loại và kiểm tra xem mô hình nào hoạt động tốt nhất. Việc thực hiện điều chỉnh siêu tham số để tối đa hóa hiệu suất của mô hình cũng rất quan trọng.

1. Logistic Regression

Text

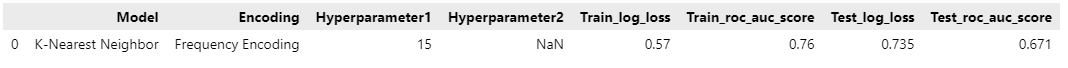
Description automatically generatedGraphical user interface, text, application

Description automatically generated

1. K-Nearest Neighbor

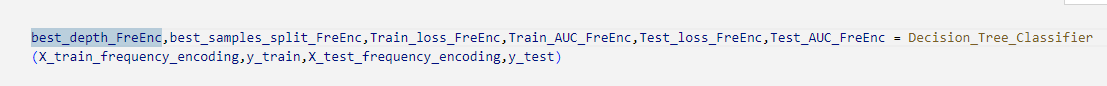
Text, letter

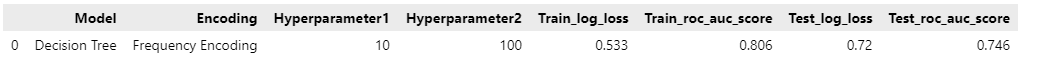
Description automatically generatedA picture containing text

Description automatically generated

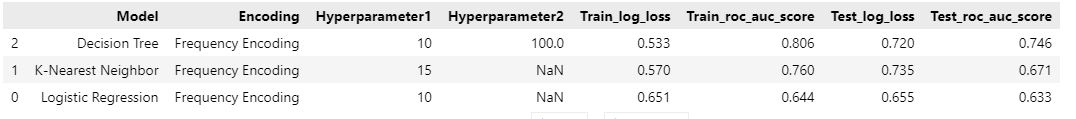
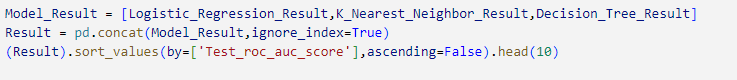
1. Decision Tree

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated with low confidence

1. So Sánh

Scatter chart

Description automatically generated