# I. Tổng quan dữ liệu

## 1. Tổng quan về 3 tệp tin:

* 3 tệp tin chứa dữ liệu về một cửa hàng bán đồ điện tử trực tuyến. Cửa hàng có trữ đồ tại 3 kho hàng nơi mà sẽ giao đồ được đặt tới khách hàng.
* Tệp tin “dirty\_data.csv” chứa các thông về người đặt hàng, mã đơn hàng, xử lý đơn hàng, thời gian, ngày tháng, địa điểm,...
* Tệp tin “missing\_data.csv” có một vài dữ liệu bị khuyết cần phải xử lý.
* Tệp tin “warehouses.csv” chứa dữ liệu về tọa độ của 3 kho hàng.

## 2. Các biến chính trong bộ dữ liệu:

* Order\_price: Giá gốc của đơn hàng (USD)
* Delivery\_charges: Phí giao hàng (USD)
* Coupon\_discount: Giảm giá trên đơn hàng (%)
* Order\_total: Thành tiền (USD)
* Season: Mùa
* Is\_expedited\_delivery: Thể hiện xem khách hàng có yêu cầu rằng đơn hàng này thuộc nhóm giao hàng nhanh hay không? (True/False)
* Distance\_to\_nearest\_warehouse: Khoảng cách từ khách hàng tới kho gần nhất (km)
* Is\_happy\_customer: Thể hiện xem khách hàng có hài lòng với dịch vụ hay không? (True/False)

## 3. Các bước thực hiện:

* Đọc dữ liệu (Import data)
* Làm sạch dữ liệu (Cleaning data); Xóa bỏ NA (dữ liệu khuyết)
* Làm rõ dữ liệu (Data visualization)
* Chuyển đổi biến (nếu có)
* Thống kê mô tả
* Thống kê suy diễn
* Xây dựng mô hình hồi quy logistic để đánh giá các nhân tố ảnh hưởng đến việc khách hàng hài lòng hay không
* Dự báo khách hàng có hài lòng với đơn hàng hay không?
* Mở rộng

# II. Kiến thức nền tảng:

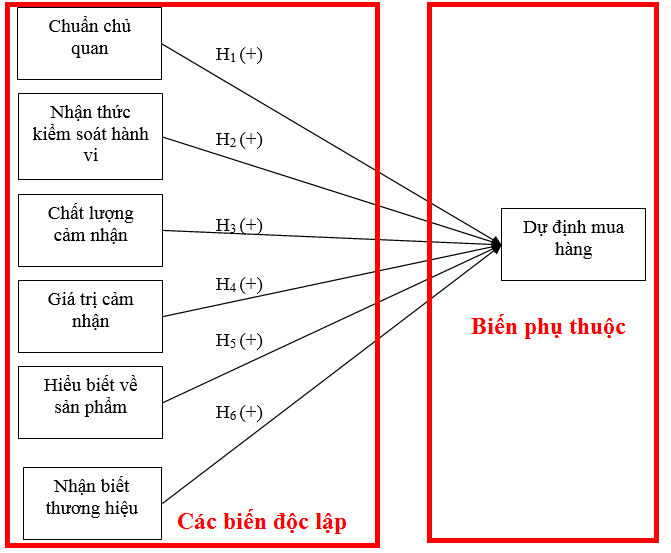
## 1. Phân tích hồi quy:

### 1.1. Định nghĩa:

Phân tích hồi quy là một kỹ thuật thống kê nhằm phân tích mối liên hệ phụ thuộc giữa một biến gọi là biến phụ thuộc (biến được giải thích, biến nội sinh) với một hoặc nhiều biến khác gọi là biến độc lập (biến giải thích, biến ngoại sinh, biến hồi quy). Mục tiêu của hồi quy là tìm ra một hàm số thể hiện mối quan hệ này, từ đó có thể dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập.

Phân tích hồi quy được sử dụng trong một số bối cảnh trong kinh doanh, tài chính và kinh tế. Về kinh tế, nó được sử dụng để giúp các nhà quản lý đầu tư định giá tài sản và hiểu mối quan hệ giữa các yếu tố như giá cả hàng hóa và cổ phiếu của các doanh nghiệp kinh doanh những mặt hàng đó.

Ví dụ: “dự định mua điện thoại hãng A”. Việc “dự định mua điện thoại” là biến phụ thuộc. Các yếu tố độc lập gồm: chuẩn chủ quan, nhận thức kiểm soát hành vi, chất lượng cảm nhận, giá trị cảm nhận, hiểu biết về sản phẩm, nhận biết thương hiệu.



### 1.2. Một số khái niệm cơ bản:

#### 1.2.1. Các thành phần của hồi quy:

* **Biến phụ thuộc (dependent variable):**

Biến phụ thuộc là biến chịu ảnh hưởng của biến khác trong một mô hình (cần dự đoán hoặc mô hình hóa). Thường là kết quả của việc biến độc lập thay đổi.

Ví dụ, nhu cầu về một hang hoá bị ảnh hưởng bởi giá cả của nó.

* **Biến độc lập (independent variable / regressor(s)):**

Biến độc lập là biến tác động tới biến khác trong một mô hình (được sử dụng để dự đoán giá trị của biến phụ thuộc). Biến này được coi là nguyên nhân hoặc yếu tố tác động đến biến phụ thuộc.

Chẳng hạn, giá hàng hoá là biến số độc lập ảnh hưởng tới lượng cầu về nó.

* **Tham số hồi quy:**

Là các tham số trong mô hình hồi quy, biểu thị mức độ và hướng của mối quan hệ giữa biến phụ thuộc và các biến độc lập.

Chẳng hạn, nếu hệ số hồi quy dương, thì khi biến độc lập tăng, biến phụ thuộc cũng có xu hướng tăng.

* **Sai số:**

Là phần chênh lệch giữa giá trị thực tế của biến phụ thuộc so với giá trị dự đoán từ mô hình hồi quy.

Sai số cho thấy mức độ chính xác của mô hình.

#### 1.2.2. Các loại hồi quy cơ bản:

* **Hồi quy đơn:**

Trong đó: : Biến phụ thuộc; : Hệ số chặn; : Sai số; : Hệ số hồi quy, biểu thị độ dốc của đường hồi quy; : Biến độc lập

* **Hồi quy bội:**

Trong đó: : Biến phụ thuộc; : Sai số; là các hệ số hồi quy và là các biến độc lập

## 2. Hồi quy logistic và logistic nhị phân:

### 2.1. Hồi quy logistic:

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.

Ví dụ: giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của họ hay không. Phân tích hồi quy logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng, trước đây, nếu khách truy cập dành hơn năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thanh toán. Nhờ vào thông tin này, sau đó, hàm hồi quy logistic có thể dự đoán hành vi của một khách mới truy cập trang web.

### 2.2. Hồi quy Binary logistic:

#### 2.2.1. Lý thuyết:

Hồi quy Binary logistic là mô hình phổ biến trong nghiên cứu dùng để ước lượng xác suất một sự kiện sẽ xảy ra. Đặc trưng của hồi quy nhị phân là biến phụ thuộc chỉ có hai giá trị: 0 và 1. Trên thực tế, có rất nhiều hiện tượng tự nhiên, hiện tượng kinh tế, xã hội,… mà chúng ta cần dự đoán khả năng xảy ra của nó như chiến dịch quảng cáo có được chấp nhận hay không, người vay có trả được nợ hay không, công ty có phá sản hay không, khách hàng có mua hay không,… Những biến nghiên cứu có hai biểu hiện như vậy được mã hóa thành hai giá trị 0 và 1, được gọi là biến nhị phân. Khi biến phụ thuộc ở dạng nhị phân, chúng ta không thể phân tích với dạng hồi quy tuyến tính thông thường vì mô hình sẽ vi phạm các giả định hồi quy. Các giả định quan trọng này bị vi phạm sẽ làm mất hiệu lực thống kê của các kiểm định trong hồi quy, dẫn đến kết quả ước lượng không còn chính xác. Trong khi đó, hồi quy Binary logistic lại không cần thiết phải thỏa mãn các giả định này.

#### 2.2.2. Phương trình:

Thay vì chúng ta ước lượng giá trị của biến phụ thuộc Y theo biến độc lập X như ở hồi quy đa biến, thì trong hồi quy Binary logistic, chúng ta sẽ ước lượng xác suất xảy ra sự kiện Y(probability) khi biết giá trị X. Biến phụ thuộc Y có hai giá trị 0 và 1, với 0 là không xảy ra sự kiện và 1 là xảy ra sự kiện. Từ đặc điểm này, chúng ta có thể đánh giá được khả năng xảy ra sự kiện (Y=1) nếu xác suất dự đoán lớn hơn 0.5, ngược lại, khả năng không xảy ra sự kiện (Y = 0) nếu xác suất dự đoán nhỏ hơn 0.5. Ta có hàm xác suất như sau: A mathematical equation with numbers and symbols

Description automatically generated

Trong đó  là xác suất xảy ra sự kiện. Thực hiện các phép chuyển đổi toán học, ta thu được hàm hồi quy mới như sau:

A black and white math symbols

Description automatically generated

Trong đó: : xác suất xảy ra sự kiện (Y=1); 1 −  xác suất xảy ra sự kiện (Y=0)

Lí do hàm hồi quy logistic có dạng như trên là vì giá trị xác suất bị ràng buộc trong phạm vi từ 0 đến 1. Trong khi đó, hàm hồi quy tuyến tính có giá trị từ âm vô cùng đến dương vô cùng. Và đối với các giá trị của hàm nằm ngoài khoảng (0,1) thì sự diễn giải theo mô hình hồi quy logistic không mang ý nghĩa thống kê.

Để giải quyết vấn đề này, có hai bước tiếp cận thông qua chúng ta thực hiện hai biến đổi:

* Thứ nhất, chúng ta sử dụng một chỉ số thống kê quan trọng đó là Odds:

A number of numbers and symbols

Description automatically generated with medium confidence

Trong đó, Odd được định nghĩa là tỉ số của hai xác suất. Nếu p là xác suất mắc bệnh, thì (1 - p) là xác suất sự kiện không mắc bệnh. Như vậy, giá trị của Odd có thể lớn hơn 1 và khi xác suất P tiến dần tới 1 thì đồng thời Odd cũng tiến ra vô cùng. Giá trị của Odd bây giờ nằm trong khoảng từ 0 đến vô cùng, vì vậy mô hình hồi quy vẫn còn hạn chế vì Odd bị giới hạn bởi 0.

Để khắc phục điều này, ta thực hiện phép biến đối tiếp theo, trình bày công thức của Odd ở trên theo logarit của Odd:



Từ đây việc hoàn chuyển từ P sang Odd và từ Odd sang log(Odd) làm cho vế phải của phương trình có dạng của hàm tuyến tính đa biến và giá trị của log(Odd) không bị giới hạn bởi cận trên, cận dưới nên việc phân tích các tham số dựa vào tích chất của hàm hồi quy đa biến sẽ dễ dàng hơn.

## 3. Phân tích phương sai (Anova):

“**ANalysis Of VAriance** (**ANOVA**)” (Phân tích phương sai), là một phương pháp thống kê dùng để so sánh sự khác biệt giữa các nhóm hoặc tập hợp dữ liệu. Dựa trên các trị trung bình của các mẫu quan sát từ các nhóm và thông qua kiểm định giả thuyết của kết luận về sự bằng nhau của các trung bình tổng thể này

### 3.1. Phân tích phương sai 1 mẫu:

#### 3.1.1. Tổng quan:

Là kiểm định sự khác biệt của trung bình giữa các nhóm dựa trên một yếu tố duy nhất.

Các giả định mô hình:

* Tổng thể có phân phối chuẩn.
* Tổng thể có phương sai bằng nhau.
* Mẫu được chọn là ngẫu nhiên và độc lập.
* Giả sử trung bình của tổng thể là thì
* Giả thuyết : “Không có sự khác biệt về trung bình giữa các nhóm” :

H0:

* Đối thuyết : “Có tồn tại sự khác biệt giữa các nhóm “

H1:

#### 3.1.2. Các bước thực hiện:

* Xét sự biến thiên giữa các nhóm để kiểm tra sự bằng nhau của trung bình giữa các nhóm:

Trong đó:

: Tổng bình phương toàn phần

SSW: Tổng bình phương bên trong các nhóm

: Tổng bình phương giữa các nhóm

* Trung bình bình phương toàn phần :
* Trung bình bình phương trong từng nhóm :
* Trung bình bình phương giữa các nhóm :
* Thống kê kiểm định :
* Bảng Anova 1 nhân tố:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nguồn của sự biến thiên | SS |  |  |  |
| Giữa các nhóm | SSB |  |  |  |
| Trong từng nhóm | SSW |  |  |
| Tổng | SST |  |  |

* Bác bỏ khi:

### 3.2. Phân tích phương sai 2 mẫu (Anova 2 nhân tố):

#### 3.2.1. Tổng quan:

ANOVA 2 nhân tố dùng để nghiên cứu tác động :

* 2 nhân tố được quan tam trên 1 biến phụ thuộc
* Tương tác giữa các mức khác nhau của 2 nhân tố
* Các giả định mô hình :
  + Tổng thể có phân phối chuẩn
  + Tổng thể có phương sai bằng nhau
  + Mẫu ngẫu nhiên được chọn độc lập

#### 3.2.2. Các bước thực hiện:

Tính tổng các bình phương:

Trong đó :

* Tổng bình phương:
  + Toàn phần :
  + Giữa các nhóm :
  + Giữa các khối:
  + Sai số :
* Các trung bình bình phương:
* Đối với các nhóm nhân tố A:

Thống kê kiểm định :

Bác bỏ khi:

* Đối với các nhóm nhân tố B:

Thống kê kiểm định :

Bác bỏ khi: