<https://www2.1010data.com/documentationcenter/beta/Tutorials/MachineLearningExamples/index_frames.html?q=BankMarketingDataSet.html>

<https://rstudio-pubs-static.s3.amazonaws.com/347011_c995dede7ed84fa78af6337be856e543.html>

Pandas visualization: <https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html>

Correlation in data analysis:

<https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec>

kaggle visualize:

<https://www.kaggle.com/learn/data-visualization>

data visualization with python coursera

<https://www.coursera.org/learn/python-for-data-visualization>

Dữ liệu có mối tương quan:

* cùng nhóm:
  + poutcome và y

0.8577140169332079 poutcome=failure ⇒ y= no

0.9116778674465034 poutcome=nonexistent ⇒ y=no

0.651128914785142 poutcome=success ⇒ y=yes

* + job và education

admin.

high.school 0.3194204567261562

university.degree 0.5520053732488965

housemaid

basic.4y 0.44716981132075473

**blue-collar**

basic.9y 0.39150637562135293

basic.4y 0.2504862762048844

**management**

university.degree 0.7055403556771546

**unknown**

unknown 0.396969696969697

**technician**

professional.course 0.4923624499480943

university.degree 0.2682782144446092

**unemployed**

university.degree 0.2583826429980276

high.school 0.25542406311637084

**retired**

basic.4y 0.3470930232558139

**self-employed**

university.degree 0.5383532723434201

**entrepreneur**

university.degree 0.41895604395604397

**student**

high.school 0.408

**services**

high.school 0.6757369614512472

* + education và marital
* khác nhóm:
  + default và age
  + contact và default

⇒ default=yes chỉ có 3 cái, còn lại đều là no hoặc unknown thì tương quan giề?

* + duration và y (do nếu duration=0 thì y=fail)
  + Các yếu tố kinh tế xã hội (emp\_var\_rate, cons\_price\_idx, cons\_conf\_idx, euribor3m, nr.employed) liên quan trực tiếp đến thời điểm xảy ra cuộc gọi cuối cùng vì các giá trị này được lấy theo thời gian xảy ra cuộc gọi cuối
* Không thể dự đoán default vì chỉ có dữ liệu của no và unknown
* các chỉ số kinh tế xã hội lấy theo báo cáo có sắn theo thời điểm của cuộc gọi cuối cùng
* housing và loan có thể dự báo dựa trên các chỉ số kinh tế xã hội nhưng không đảm bảo vì các thông số này của mọi người là giống nhau trong từng giai đoạn
* dự đoán y cần tất cả thuộc tính để đạt được độ chính xác tốt
* month, day\_of\_week, campaign, pdays, previous, poutcome được ghi nhận khi đã xảy ra và là các kết quả hoàn toàn ngẫu nhiên

**CPI**

Để tính toán chỉ số giá tiêu dùng người ta tính [số bình quân gia quyền](https://vi.wikipedia.org/wiki/S%E1%BB%91_b%C3%ACnh_qu%C3%A2n_gia_quy%E1%BB%81n) theo [công thức Laspeyres](https://vi.wikipedia.org/w/index.php?title=C%C3%B4ng_th%E1%BB%A9c_Laspeyres&action=edit&redlink=1) của giá cả của kỳ báo cáo (kỳ *t*) so với kỳ cơ sở. Để làm được điều đó phải tiến hành như sau:

1. Cố định [giỏ hàng hóa](https://vi.wikipedia.org/wiki/Gi%E1%BB%8F_h%C3%A0ng_h%C3%B3a): thông qua điều tra, người ta sẽ xác định lượng hàng hoá, dịch vụ tiêu biểu mà một người tiêu dùng điển hình mua.

2. Xác định giá cả: thống kê giá cả của mỗi mặt hàng trong giỏ hàng hoá tại mỗi thời điểm.

3. Tính chi phí (bằng tiền) để mua giỏ hàng hoá bằng cách dùng số lượng nhân với giá cả của từng loại hàng hoá rồi cộng lại.

4. Lựa chọn thời kỳ gốc để làm cơ sở so sánh rồi tính chỉ số giá tiêu dùng bằng công thức sau:

|  |  |
| --- | --- |
| CPIt = 100 x | Chi phí để mua giỏ hàng hoá thời kỳ *t* |
| Chi phí để mua giỏ hàng hoá kỳ cơ sở |

Thời kỳ gốc sẽ được thay đổi trong vòng 5 đến 7 năm tùy ở từng nước.

**CCI**

Tại Hoa Kỳ, chỉ số niềm tin người tiêu dùng Hoa Kỳ (CCI) là một chỉ số kinh tế để đo lường lòng tin của người tiêu dùng, định nghĩa mức độ lạc quan về tình trạng của nền kinh tế Hoa Kỳ của người tiêu dùng thể hiện thông qua các hoạt động tiết kiệm và chi tiêu của họ.

Nói một cách dễ hiểu, niềm tin của người tiêu dùng tăng lên cho thấy sự tăng trưởng kinh tế trong đó người tiêu dùng đang chi tiền, cho thấy mức tiêu dùng cao hơn. Niềm tin của người tiêu dùng giảm đồng nghĩa với việc tăng trưởng kinh tế chậm lại và do đó người tiêu dùng có xu hướng giảm chi tiêu. Ý tưởng cho rằng những người càng cảm thấy tự tin hơn về nền kinh tế cũng như công việc và thu nhập của họ, thì họ càng có nhiều khả năng mua hàng. Niềm tin của người tiêu dùng suy giảm là một dấu hiệu của tăng trưởng kinh tế chậm lại và có thể cho thấy nền kinh tế đang gặp khó khăn.

Mỗi tháng Ban Hội nghị khảo sát 5.000 hộ gia đình Hoa Kỳ. Cuộc khảo sát bao gồm năm câu hỏi hỏi ý kiến ​​của người trả lời về những điều sau:

1. Điều kiện kinh doanh hiện tại
2. Điều kiện kinh doanh trong sáu tháng tới
3. Điều kiện việc làm hiện tại
4. Điều kiện việc làm trong sáu tháng tới
5. Tổng thu nhập của gia đình trong sáu tháng tới

Những người tham gia khảo sát được yêu cầu trả lời từng câu hỏi là "tích cực", "tiêu cực" hoặc "trung lập". Kết quả sơ bộ từ cuộc khảo sát niềm tin của người tiêu dùng được công bố vào thứ Ba cuối cùng của mỗi tháng lúc 10 giờ sáng theo giờ EST.

Khi dữ liệu đã được thu thập, một tỷ lệ được gọi là "giá trị tương đối" được tính toán riêng cho từng câu hỏi. Các câu trả lời tích cực của mỗi câu hỏi được chia cho tổng các câu trả lời tích cực và tiêu cực. Giá trị tương đối cho mỗi câu hỏi sau đó được so sánh với mỗi giá trị tương đối từ năm 1985. Việc so sánh các giá trị tương đối này dẫn đến "giá trị chỉ số" cho mỗi câu hỏi.

Các giá trị chỉ số cho cả năm câu hỏi sau đó được tính trung bình với nhau để tạo thành chỉ số niềm tin của người tiêu dùng; trung bình của các giá trị chỉ số cho câu hỏi một và ba tạo thành chỉ số tình hình hiện tại và trung bình của các giá trị chỉ số cho câu hỏi hai, bốn và năm hình thành chỉ số kỳ vọng. Dữ liệu được tính cho toàn bộ Hoa Kỳ và cho từng khu vực trong số chín khu vực điều tra dân số của đất nước.

**Ở Bồ Đào Nha**, chỉ số niềm tin của người tiêu dùng trên ba kỳ hạn trung bình động dựa trên các cuộc phỏng vấn với người tiêu dùng về nhận thức của họ về tình hình kinh tế hiện tại và tương lai của đất nước và xu hướng mua hàng của họ. Nó được ước tính bằng cách sử dụng sự khác biệt giữa tỷ lệ các phản hồi đánh giá tích cực và các phản hồi đánh giá tiêu cực, nhưng không bao gồm tỷ lệ các phản hồi trung lập. Việc sử dụng các đường trung bình động giúp làm mượt chuỗi bằng cách loại bỏ các chuyển động bất thường, cho phép phát hiện các xu hướng ngắn hạn.

Chỉ số niềm tin của người tiêu dùng là giá trị trung bình số học của các số dư (tính bằng điểm phần trăm) của các câu trả lời cho các câu hỏi cho người tiêu dùng về kỳ vọng của họ trong 12 tháng tới liên quan đến tình hình tài chính của hộ gia đình họ, tình hình kinh tế chung, kỳ vọng thất nghiệp và tiết kiệm hộ gia đình.

**Euribor (Euro InterBank Offered Rate)** là lãi suất trung bình mà tại đó một số ngân hàng cung cấp cho nhau các khoản vay ngắn hạn bằng đồng euro.

Lãi suất Euribor 3 tháng là lãi suất mà một số ngân hàng châu Âu cho vay lẫn nhau bằng đồng euro, theo đó các khoản vay có thời gian đáo hạn là 3 tháng.

Khi lãi suất Euribor tăng hoặc giảm (về cơ bản), rất có thể lãi suất của các sản phẩm ngân hàng như thế chấp, tài khoản tiết kiệm và cho vay cũng sẽ được điều chỉnh.

1) Do sự thay đổi chính sách của Viện Thị trường Tiền tệ Châu Âu, tỷ giá Euribor chỉ được cung cấp cho công chúng với thời gian trì hoãn 24 giờ. Kể từ tháng 3 năm 2014, không có trang web nào được phép công khai tỷ lệ thời gian thực nữa.

2) Kể từ ngày 1 tháng 12 năm 2018, số lượng giá cước Euribor đã giảm xuống còn 5 (1 tuần, 1, 3, 6 và 12 tháng).

Cách tính:

Một hội đồng đại diện của các ngân hàng cung cấp báo giá hàng ngày của tỷ lệ, được làm tròn đến hai chữ số thập phân, mà mỗi Ngân hàng Hội nghị tin rằng một ngân hàng chính đang báo giá cho một ngân hàng chính khác đối với các khoản tiền gửi có kỳ hạn liên ngân hàng trong khu vực đồng Euro, cho kỳ hạn từ một tuần đến một năm . Mọi Ngân hàng Panel được yêu cầu nhập trực tiếp dữ liệu của mình không muộn hơn 11:00 sáng (CET) vào mỗi ngày khi hệ thống Chuyển khoản nhanh thanh toán gộp tự động xuyên châu Âu (TARGET) mở. Vào lúc 11:02 sáng (CET), GRSS (Hệ thống đặt giá toàn cầu) sẽ ngay lập tức công bố tỷ giá tham chiếu trên Refinitiv (ví dụ: Reuters), Bloomberg và một số nhà cung cấp thông tin khác, sau đó sẽ được cung cấp cho tất cả người đăng ký của họ. Tỷ lệ được công bố là giá trị trung bình được làm tròn, cắt ngắn của các tỷ lệ được trích dẫn: 15% báo giá cao nhất và thấp nhất bị loại bỏ, phần còn lại được tính trung bình và kết quả được làm tròn đến 3 chữ số thập phân. Tỷ giá Euribor là tỷ giá giao ngay (Tỷ giá giao ngay là lãi suất cho một kỳ hạn cụ thể, được sử dụng để chiết khấu các dòng tiền xảy ra tại ngày đó, thường có hiệu lực trong 2 ngày làm việc kể từ khi giao dịch)

|  |  |
| --- | --- |
| Numeric data | Catagorical data |
| Age  Duration  Campaign (số lượng liên hệ được thực hiện trong campaign này và cho khách hàng này, bao gồm cả liên hệ cuối cùng)  Pdays (999 means client was not previously contacted)  Cons.price.idx (chỉ số giá tiêu dùng: tỉ lệ giá tiền mua nhu yếu phẩm so với thời kỳ cơ sở)  <https://tradingeconomics.com/portugal/consumer-price-index-cpi>  Cons.conf.idx (chỉ số niềm tin tiêu dùng, trong data chỉ có số âm trong khi bảng của US>0 ???) consumer confidence != consumer confidence index  <https://tradingeconomics.com/portugal/consumer-confidence>  <https://data.oecd.org/leadind/consumer-confidence-index-cci.htm>  <https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&userLoadSave=Load&userTableOrder=4&tipoSeleccao=1&contexto=pq&selTab=tab1&submitLoad=true>  <https://www.economy.com/portugal/consumer-confidence/not-seasonally-adjusted>  Emp.var.rate  <https://tradingeconomics.com/portugal/employment-change>  Euribor3m  Nr.employed | Job  Marital ( "divorced" có nghĩa là đã ly hôn hoặc góa bụa)  Education (order)  Default (khả năng vỡ nỡ)  Housing  Loan  Contact  Month  Previous (số lượng liên hệ được thực hiện trước chiến dịch này và cho khách hàng này)  Day\_of\_week  Poutcome (kết quả của đợt marketing trước)  y |

**Paper 103-2012 - CSI: Customer Segmentation Intelligence for Increasing Profits**

**Darius Baer, Ph.D., SAS Institute Inc., Cary, NC**

* có nhiều phương pháp segmentation tuy nhiên cái nào hoạt động thì phải dựa trên mục tiêu kinh doanh và thuộc tính của khách hàng.
* customer relationship management (CRM): cung cấp các mối liên hệ cá nhân với từng khách hàng, tối ưu hóa sự hài lòng nhằm tối ưu hóa lợi nhuận, có thể xem như tiếp thị 1:1.

Phân đoạn ở đây được thực hiện bằng cách sử dụng business rule, statistical clustering (phân nhóm thống kê) hoặc kết hợp cả hai.

Có 4 phương pháp nhóm/phân đoạn chính:

1. Business rule (không bao gồm việc sử dụng clustering): có xác định trước các thuộc tính được sử dụng, chỉ định mỗi khách hàng vào một nhóm theo các lớp đã được định trước (ngược với phân cụm trong đó dùng một thuật toán thống kê để xác định nhóm theo các phép đo mức độ tương tự của thuộc tính)

* Phân nhóm dựa trên demography (độ tuổi, giới tính, thu nhập…):

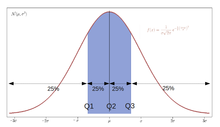
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| younger female | middle female | older female | younger male | middle male | older male |
| F  <25 | F  <55 | F  >=55 | M  <25 | M  <55 | M  >=55 |

1. Quantile membership: Tạo phân đoạn dựa trên thành viên lượng tử sử dụng tần số thuộc tính để xác định thành viên phân đoạn(không dùng clustering nhưng grouping bằng cách đánh giá các thống kê đơn biến (dữ liệu chỉ bao gồm các quan sát về một thuộc tính, ví dụ tiền lương của công nhân trong ngành) về các thuộc tính của khách hàng)

RFM: Áp dụng các quy tắc tạo nhóm(khách mua sắm trong 2 tháng qua và đã mua 20 lần trong năm qua và chi trung bình từ 500 usd mỗi tuần ⇒gold segment)

sử dụng thống kê đơn biến để tạo các tập lượng tử (top 10% khách của RFM thuộc gold segment)

Dùng thuộc tính RFM để phân đoạn bằng quantile membership, nên chia các metric thành các quantile (các chặn trong thống kê) (hợp lý là 5: 0 20 40 60 80 100)



Gán các giá trị từ A - E cho từng phân vị của RFM. Từ đây chúng ta sẽ có 125 phân phúc. Vì có quá nhiều phân khúc cần quản lý, ta gộp lại như sau:

A/B trong cả 3 RFM → khách hàng rất có giá trị

D/E trong cả 3 RFM → khách hàng có giá trị rất thấp

Và một số phân khúc khác tùy theo mục đích kinh doanh và hướng phát triển

Ưu:

phù hợp mục tiêu kinh doanh

Dễ áp dụng

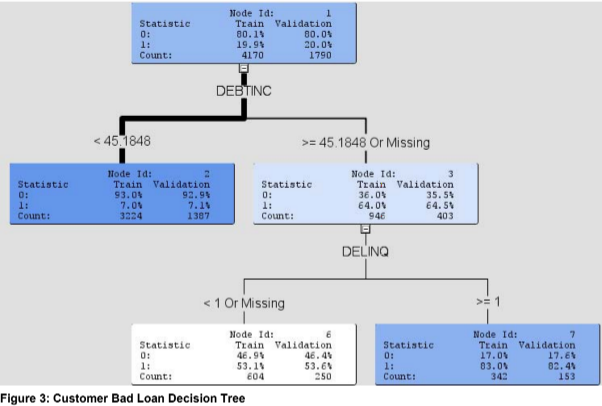
Khuyết:

có thể không phản ảnh được hành vi thực tế của khách

một số phân đoạn không có khác biệt đáng kể với nhau

1. Phân cụm có giám sát (ví dụ decision tree): Cần một target cụ thể

cách này áp đặt thành kiến trong quá trình phỏng đoán có học thức khách hàng có khả năng mua nhất.



Phân đoạn khả năng vỡ nợ dựa trên hai thuộc tính debtinc (tỉ lệ nợ trên thu nhập - giá trị cao cho thấy khả năng mắc nợ lớn hơn) và delinq(số lần nợ thế chấp trong 3 năm qua -- giá trị khác 0 là xấu)

Lựa chọn giá trị từng phân đoạn là một vấn đề khó tiếp cận, nhất là đối với giá trị thực.

Ưu: Có thể tương quan mục tiêu với các thuộc tính của khách hàng

Khuyết: Giả định các thuộc tính ảnh hưởng đến target không đổi trong tương lai

1. Phân cụm không giám sát (không được định hướng bởi các giá trị của các biến cụ thể)

Dùng khoảng cách Euclide đo lường mức độ giống nhau của từng thuộc tính khách hàng cùng với kmeans

“Doughnut Method for Cluster Formation” (chuyển đổi giá trị của biến hoặc thuộc tính thành Z-score (khoảng cách đến điểm dữ liệu trung bình: **z = (x – μ) / σ**) dựa trên giá trị trung bình và độ lệch chuẩn của từng biến độc lập, cho phép mỗi biến hoặc thứ nguyên có nhiều đóng góp tương tự hơn cho sự hình thành cụm. Giải quyết được vấn đề center bị chia thành nhiều nhóm nhỏ không cần thiết.

Doughnut gom nhóm center trước, các cụm khác sẽ được phân chia một cách thích hợp. Trong thực tế, độ lệch chuẩn của kích thước cụm trung tâm sẽ nhỏ hơn các cụm bên ngoài.

* tính Z-score dữ liệu
* đánh trọng số các dữ liệu quan trọng
* Giảm chiều dữ liệu (PCA)
* Xóa bỏ khách hàng chỉ mua 1 lần duy nhất. Tính khoảng cách mỗi khách hàng đến mean tổng thể. Bỏ dữ liệu ở trung tâm. Sắp xếp từ lớn đến nhỏ dựa trên khoảng cách vừa tính. giữ lại 60% khách hàng xa mean nhất
* Bước phân cụm thứ hai từ doughnut chia cụm đúng số lượng và giữ chúng càng cách xa nhau càng tốt
* ổn định trọng tâm các cụm bằng cách bỏ các giá trị ngoại lai, đặt lại các center sau đó đưa các điểm ngoại lai trở lại
* Nhóm dữ liệu bị loại trừ trước lúc làm doughnut, gọi là cụm trung tâm
* di chuyển các điểm dữ liệu ngoại lai của trung tâm về cụm gần nhất

**CÁC PHƯƠNG PHÁP CUSTOMER SEGMENTATION**

1. Business rule (Bae 2012)

* Phương pháp cụ thể:
* Gom nhóm theo các lớp đã xác định từ trước?
* Dữ liệu bao gồm: Demographic, Purchase history
* Ưu:
  + Dễ áp dụng?
  + Dễ truy vấn cơ sở dữ liệu (tại sao?)
* Khuyết:
  + Không tập trung vào hành vi của khách hàng (do dữ liệu) ⇒ có thể không dự đoán được tình huống hiện tại và tương lai

1. Customer profiling (Colica 2011)
2. Magento (2014)
3. Quantile membership (Baer 2012)
4. RFM cell classification grouping (Colica 2011)
5. Supervised clustering with decision tree (Baer 2012)
6. Customer likeness clustering (Colica 2011)
7. Purchase affinity clustering (Colica 2011)
8. Unsupervised clustering ( Colica 2011)