Mô hình trong phân tích dữ liệu

Jan - 2024

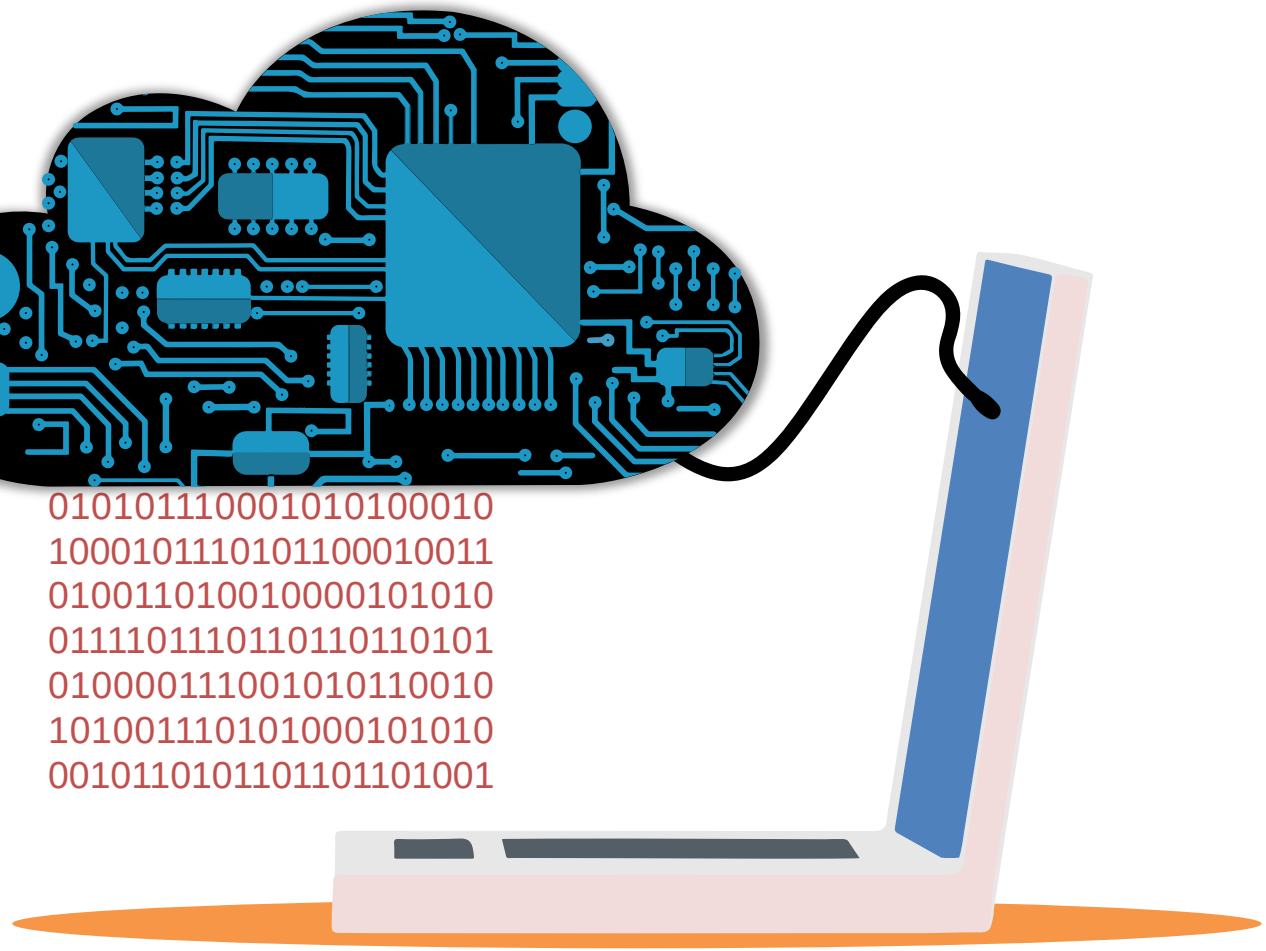
Huongdtl15@vpbank.com.vn

Nội dung

1.Tổng quan về phân tích và các mô hình trong phân tích dữ liệu

2. Mô hình thường dùng cho phòng SBD

3. Tổng kết

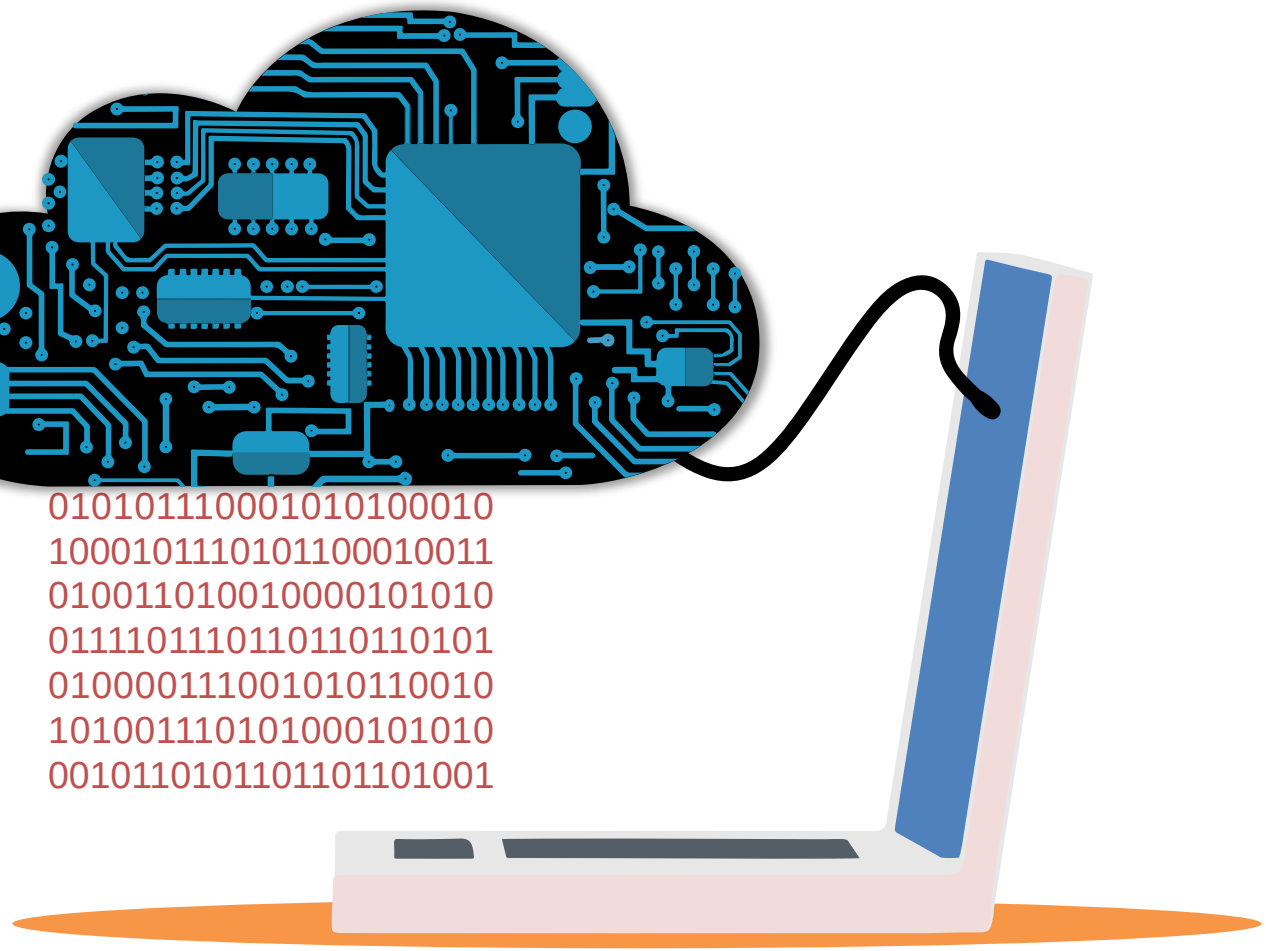


Nội dung

1.Tổng quan về phân tích và các mô hình trong phân tích dữ liệu

2. Mô hình thường dùng cho phòng SBD

3. Tổng kết



Trong phân tích dữ liệu, có 3 loại phân tích chính:

Phân tích khác biệt (Discriminant Analysis -DA)

Là một phương pháp phân tích trong thống kê được dùng rất nhiều trong Data mining để phân loại các đối tượng (object) vào các nhóm dựa trên việc đo lường các đặc trưng của đối tượng

Phân tích liên hệ (association analysis)

Là phương pháp phân tích dữ liệu để tìm ra các mối liên kết hoặc sự kết hợp giữa các sản phẩm hoặc dịch vụ trong dữ liệu, như sản phẩm nào thường được mua cùng nhau, sản phẩm nào có thể gợi ý cho khách hàng khi họ mua sản phẩm khác,..

Phân tích tương quan (correlation analysis) và tiên lượng (prediction)

Là một nhóm các kĩ thuật dùng để đo lường mức độ liên hệ giữa các biến. Phân tích tương quan bàn về đặc thù có thể đo lường trong mối liên hệ giữa các biến ở việc sự thay đổi giá trị của biến này sẽ gây nên ảnh hưởng tới sự thay đổi và phân bố xác suất của biến kia như thế nào

Với phạm vi công việc và các đề bài phân tích của phòng SBD hiện nay, thông thường sẽ sử dụng phân tích tương quan và tiên lượng.

Tùy vào loại phân tích và mục tiêu phân tích để lựa chọn mô hình phù hợp

Khái niệm “mô hình”

Mô hình là một **phương trình** để mô tả mối liên quan giữa các biến. Xây dựng mô hình chính là đi tìm quy luật (phương trình) liên quan giữa các biến với nhau.

Các loại mô hình thường dùng trong phân tích dữ liệu

Mô hình thống kê, mô hình máy học, mô hình mạng nơ-ron, **mô hình hồi quy** (tuyến tính hoặc phi tuyến)...

Phân tích tương quan và tiên lượng thường “song hành” cùng các mô hình hồi quy. Phân tích tương quan (correlation) chỉ đánh giá được mối liên hệ và hệ số tương quan giữa các biến, không phân biệt biến độc lập và biến phụ thuộc . Do đó, cần phải sử dụng thêm các mô hình hồi quy để tiên lượng mức độ ảnh hưởng của việc thay đổi giá trị của biến độc lập lên biến phụ thuộc.

Với phạm vi công việc và các yêu cầu hiện nay của phòng SBD, nên nghĩ tới việc sử dụng các mô hình hồi quy để đáp ứng các yêu cầu của clients.

Các mô hình hồi quy trong phân tích dữ liệu

Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hồi quy tuyến tính sử dụng hàm tuyến tính. (Hàm tuyến tính là các hàm số bậc 1 và thỏa mãn nguyên lý xếp chồng)

Mô hình hồi quy phi tuyến tính

Mô hình hồi quy phi tuyến tính là một dạng phân tích hồi quy trong đó dữ liệu quan sát được mô hình hóa bằng một hàm là một sự kết hợp phi tuyến tính của các tham số mô hình và phụ thuộc vào một hay nhiều biến độc lập. Các hàm phi tuyến tính gồm hàm mũ, hàm logarit, hàm lượng giác, hàm lũy thừa...

Từ mô hình hồi quy tuyến tính và phi tuyến tính, có thể chia thành rất nhiều loại mô hình hồi quy khác nhau tùy theo tính chất của biến và các điều kiện kèm theo. Ví dụ:

- Hồi quy tuyến tính: hồi quy tuyến tính đơn biến, đa biến, hồi quy tuyến tính với biến tiên lượng là biến phân nhóm...

- Hồi quy phi tuyến tính: hồi quy logistic, hồi quy logistic đa thức, hồi quy Cox...

Điều quan trọng cần ghi nhớ: hệ số tương quan và mô hình tuyến tính chỉ phản ánh mối quan hệ tương quan thống kê (ví dụ A đồng biến/nghịch biến với B ở mức độ nào...) chứ không phản ánh mối quan hệ nhân quả (A suy ra B, hoặc nếu có A sẽ dẫn tới B...)

Tổng kết phần 1

Keywords:

Phân tích tương quan (correlation analysis) và tiên lượng (prediction)...

...thường "song hành" cùng các mô hình hồi quy

Mô hình ...

...là một phương trình để mô tả mối liên quan giữa các biến.

Mô hình hồi quy ...

...tuyến tính và phi tuyến tính

Nội dung

1.Tổng quan về phân tích và mô hình trong phân tích dữ liệu

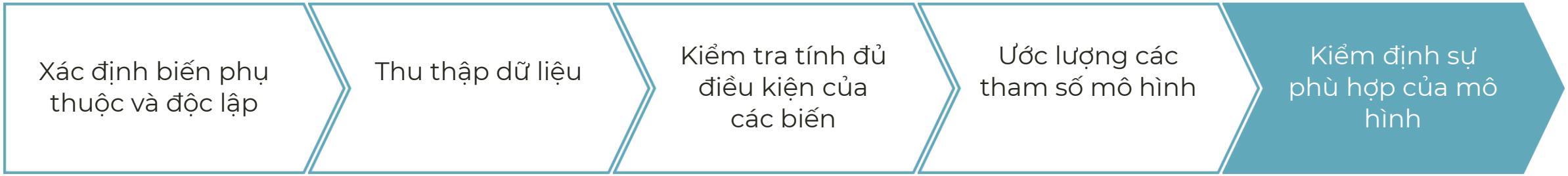
2. Mô hình thường dùng cho phòng SBD

3. Tổng kết

“"Regression is the study of dependence."”

GEORGE E. P. BOX

Các bước xây dựng mô hình



Xác định biến phụ thuộc Y và các biến độc lập X sẽ sử dụng trong mô hình

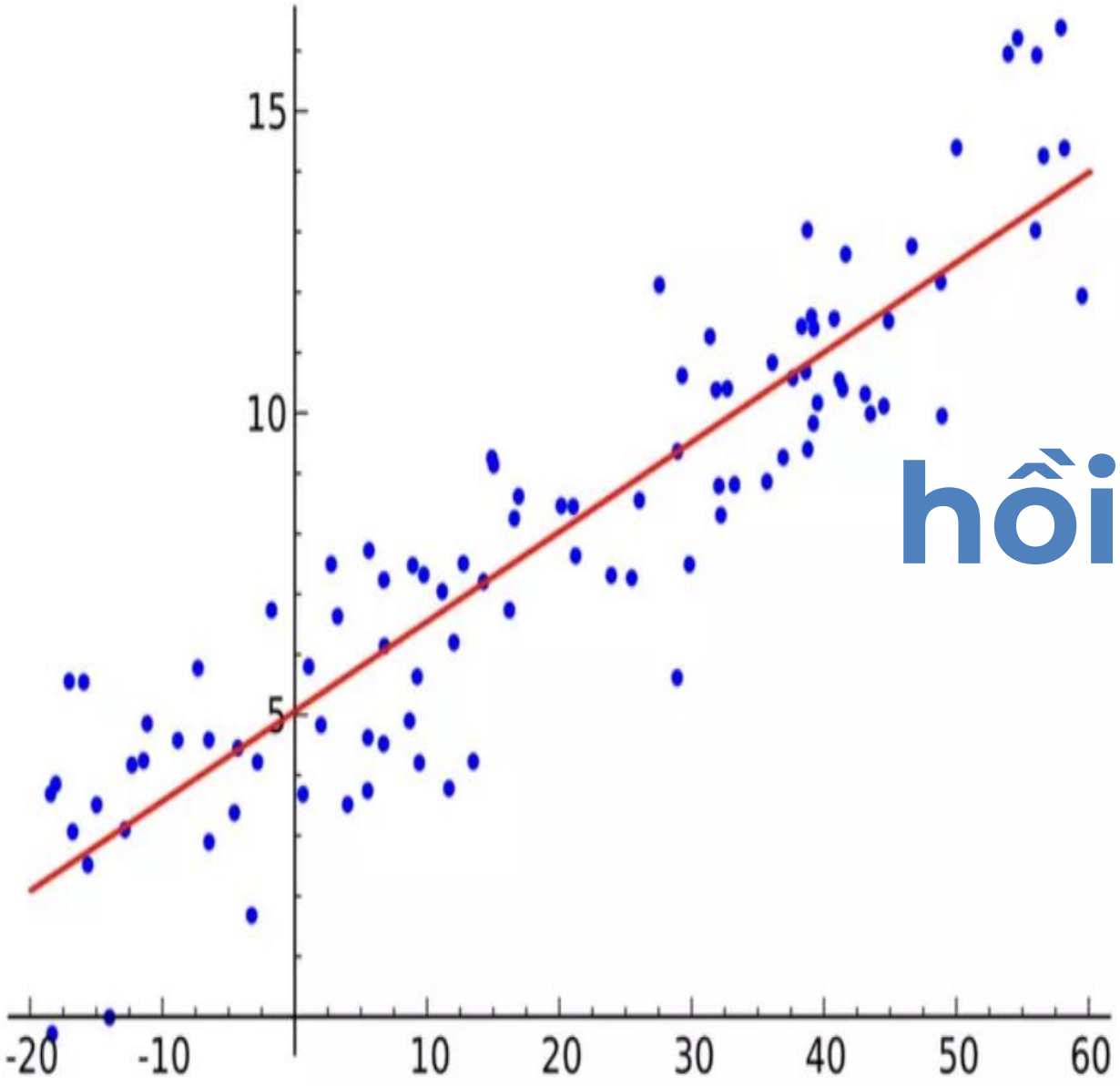
Thu thập dữ liệu cần thiết cho các biến Y và X

Kiểm tra xem các biến có đáp ứng được các giả định của mô hình hồi quy định sử dụng hay không

Sử dụng phương pháp ước lượng thích hợp để ước lượng các tham số trong mô hình

Thực hiện các kiểm định thống kê để đánh giá sự phù hợp của mô hình

Trước khi thực hiện hồi quy, nên vẽ biểu đồ phân tán để nhận dạng kiểu biến đổi dữ liệu nhằm chọn mô hình hồi quy phù hợp. Khi đó kết quả về mô hình, đánh giá, kiểm định mới đáng tin.



Mô hình hồi quy tuyến tính

Linear regression model

Mô hình hồi quy tuyến tính là gì?

- Định nghĩa: Mô hình thể hiện mối quan hệ Y phụ thuộc vào X về trung bình và ảnh hưởng của các yếu tố ngẫu nhiên
- Phương trình tổng quát: $Y = \beta_1 + \beta_2 X + u$

Trong đó:

- $\boldsymbol{u}$: Sai số ngẫu nhiên (*random error*)
- $\boldsymbol{\beta_1, \beta_2}$: Các hệ số hồi quy (*regression coefficient*)

Mô hình hồi quy tuyến tính phù hợp sử dụng với bộ dữ liệu gồm các biến liên tục (continuous data) như: tuổi tác, chiều cao, thu nhập, mức chi tiêu ...

Ví dụ:

Đề bài: Phân tích mức độ hài lòng của khách hàng khi sử dụng dịch vụ ngân hàng điện tử

Ta có:

Biến phụ thuộc (Y): Mức độ hài lòng của khách hàng.

Biến độc lập (X):

- Dễ sử dụng ứng dụng: Đánh giá về độ dễ sử dụng của ứng dụng ngân hàng điện tử.
- Tính năng: Đánh giá về tính năng của ứng dụng, ví dụ như chuyển khoản, thanh toán hóa đơn, và kiểm tra tài khoản.
- Tốc độ giao dịch: Thời gian mà giao dịch được thực hiện từ khi khách hàng yêu cầu đến khi hoàn tất.
- Hỗ trợ trực tuyến: Chất lượng hỗ trợ trực tuyến từ dịch vụ khách hàng.

Phương trình mô hình hồi quy tuyến tính có thể có dạng như sau:

$$\text{Mức độ hài lòng} = \beta_0 + \beta_1 \times \text{Dễ sử dụng ứng dụng} + \beta_2 \times \text{Tính năng} + \beta_3 \times \text{Tốc độ giao dịch} + \beta_4 \times \text{Hỗ trợ trực tuyến} + \epsilon$$

Dữ liệu biến độc lập X là các dữ liệu lịch sử thu thập được từ trước. Sau khi xử lý làm sạch dữ liệu, các công cụ như R, Python... sẽ giúp chúng ta tìm ra các hệ số hồi quy gần đúng nhất với bộ dữ liệu có sẵn.

Ưu điểm

Dễ hiểu và giải thích

Mô hình hồi quy tuyến tính dễ hiểu và giải thích mối quan hệ giữa các biến.

Phù hợp với nhiều loại dữ liệu

Mô hình có thể áp dụng với nhiều loại dữ liệu khác nhau như số liệu kinh tế, dân số,...

Có thể dự đoán

Mô hình cho phép dự đoán giá trị của biến phụ thuộc dựa trên các giá trị của biến độc lập.

Phân tích tương quan

Mô hình cho phép phân tích mối tương quan giữa các biến và mức độ ảnh hưởng của từng biến độc lập.

Đơn giản tính toán

Các phương pháp tính toán và ước lượng mô hình tương đối đơn giản.

Nhược điểm của mô hình

Giả định mối quan hệ tuyến tính

Mô hình giả định mối quan hệ giữa các biến là tuyến tính, trong khi đó thực tế có thể không phải vậy.

Nhiễu trong dữ liệu

Các sai số ngẫu nhiên trong dữ liệu có thể dẫn đến ước lượng tham số sai lệch.

Đa cộng tuyến

Sự tồn tại đa cộng tuyến giữa các biến độc lập có thể làm giảm độ chính xác của mô hình.

Khó xác định mối quan hệ nhân quả

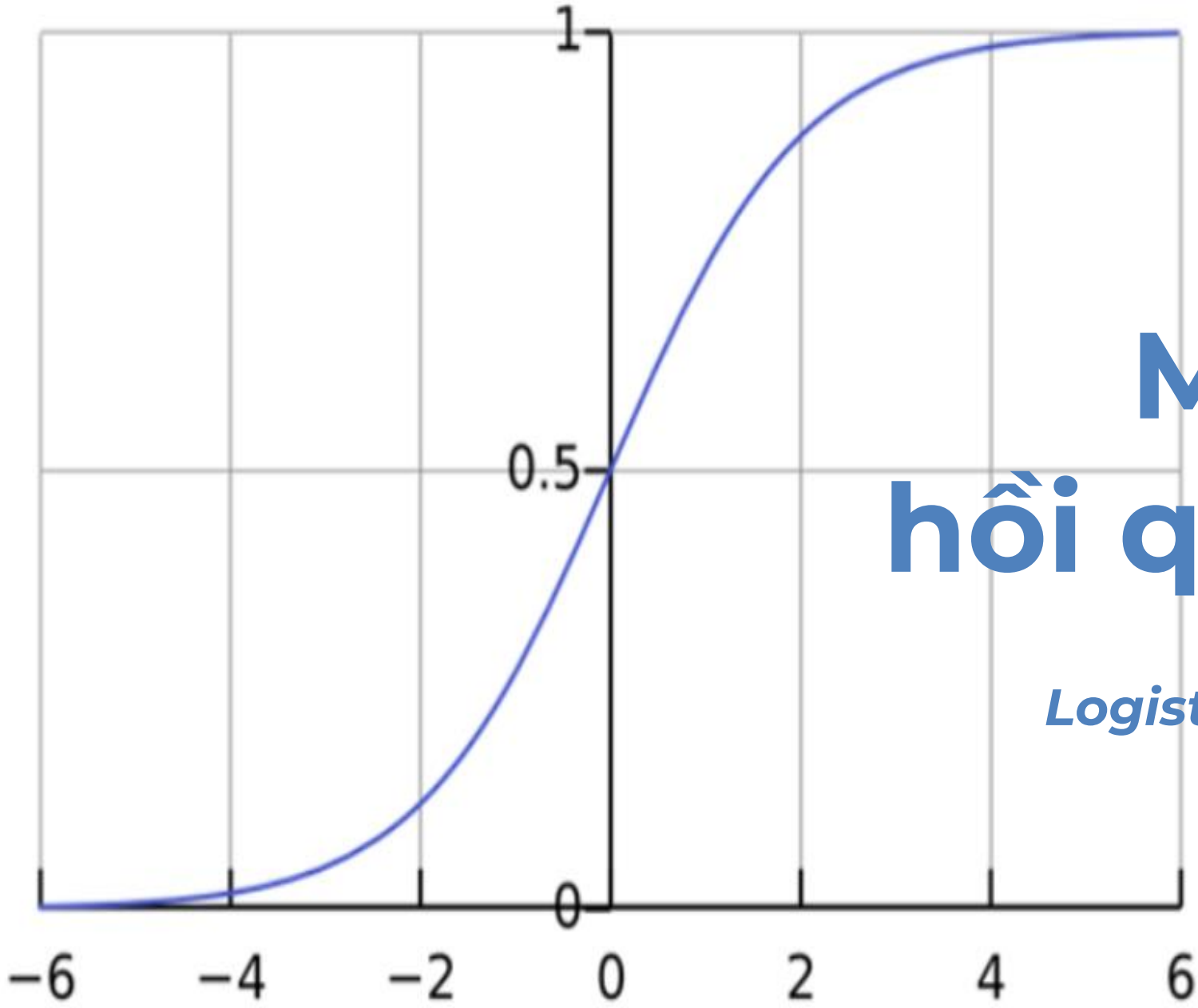
Mô hình hồi quy chỉ mô tả mối quan hệ thống kê, không thể hiện mối quan hệ nhân quả.

Phụ thuộc vào dữ liệu quan sát

Kết quả mô hình phụ thuộc hoàn toàn vào chất lượng và độ đại diện của dữ liệu.

Giả định phân phối chuẩn của sai số

Mô hình giả định sai số tuân theo phân phối chuẩn, điều này có thể không đúng trong thực tế.



Mô hình hồi quy Logistic

Logistic regression model

Mô hình hồi quy logistic là gì?

- Định nghĩa: là mô hình dự đoán xác suất của một sự kiện nhị phân xảy ra dựa trên các biến độc lập
- Phương trình tổng quát:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

Trong đó:

- $P(Y=1)$ là xác suất sự kiện Y thuộc nhóm 1.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ là các hệ số hồi quy cần được ước lượng.
- $X_1, X_2, \ldots, X_n$ là các biến độc lập.
- e là số Euler, có giá trị khoảng 2.71828.

Mô hình hồi logistic phù hợp sử dụng với bộ dữ liệu gồm các biến phân loại (binary/categorical data) như: yes/no, tốt/xấu, tích cực/tiêu cực, rời bỏ/ở lại...

Ví dụ:

Đề bài: Tính xác suất phát sinh nợ xấu của khách hàng

Ta có:

Biến phụ thuộc (Y): Khách hàng có nợ xấu (1) hoặc không có nợ xấu (0).

Biến độc lập (X):

- Điểm tín dụng: Điểm tín dụng của khách hàng theo xếp loại hệ thống.
- Tổng nợ: Tổng dư nợ của KH tại NH.
- Tỉ lệ nợ thu nhập: Tỉ lệ giữa tổng nợ và thu nhập hàng tháng.
- Số năm làm việc: Số năm mà khách hàng đã làm việc.
- Số lượng nợ tại tất cả các NH: Số lượng khoản vay tại tất cả các NH theo CIC

Phương trình mô hình hồi quy logistic có thể có dạng như sau:

$$P(\text{Nợ xấu} = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1\times\text{Điểm tín dụng}+\beta_2\times\text{Tổng nợ}+\beta_3\times\text{Tỉ lệ nợ thu nhập}+\beta_4\times\text{Số năm làm việc}+\beta_5\times\text{Số lượng tài khoản tín dụng})}}$$

Dữ liệu biến độc lập X là các dữ liệu lịch sử thu thập được từ trước. Sau khi xử lý làm sạch dữ liệu, các công cụ như R, Python... sẽ giúp chúng ta tìm ra các hệ số hồi quy gần đúng nhất với bộ dữ liệu có sẵn.

Ưu điểm

Xử lý tốt dữ liệu nhị phân

Mô hình hồi quy logistic có thể xử lý dữ liệu nhị phân, phân loại các quan sát thành 2 nhóm

Kết quả dễ hiểu

Kết quả dự báo của mô hình dưới dạng xác suất, dễ dàng hiểu và giải thích

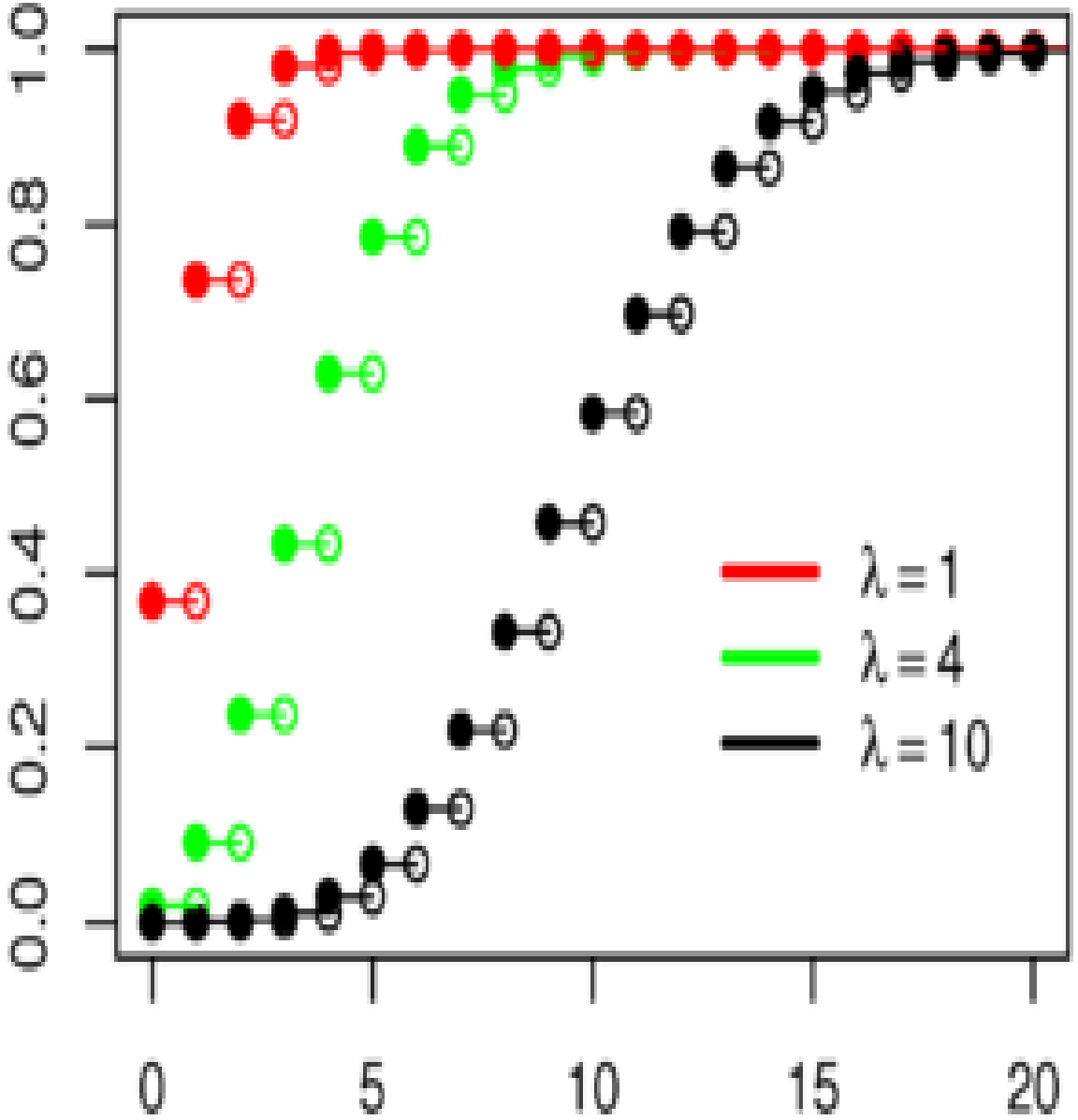
Nhược điểm

Giả định tuyến tính

Tính chất tuyến tính của mô hình hồi quy logistic không phù hợp với một số dữ liệu thực tế

Khó xử lý dữ liệu đa chiều

Mô hình gặp khó khăn trong việc xử lý các dữ liệu có nhiều chiều, cần kỹ thuật xử lý dữ liệu phù hợp



Mô hình hồi quy Poisson

Poisson regression model

Mô hình hồi quy Poisson là gì?

- Định nghĩa: là mô hình được sử dụng khi biến phụ thuộc là biến đếm (count data), tức là là số lượng các sự kiện đếm được trong một khoảng thời gian hoặc không gian cố định
- Phương trình tổng quát:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Trong đó:

- $\log(\lambda)$ là logarit tự nhiên của giá trị kỳ vọng (λ) của biến đếm.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ là các hệ số hồi quy cần được ước lượng.
- $X_1, X_2, \ldots, X_n$ là các biến độc lập.

Mô hình hồi quy Poisson phù hợp sử dụng với bộ dữ liệu gồm các biến mà bản chất là các số đếm (count), không phải là biến liên tục hay biến nhị phân như: Số người đang xếp hàng chờ ở quầy giao dịch, số lượt khách hàng được phục vụ mỗi ngày ở chi nhánh...

Ví dụ:

Đề bài: Phân tích số lượng giao dịch tài khoản ngân hàng mỗi ngày dựa trên các yếu tố liên quan.

Ta có:

Biến phụ thuộc (Y): Số lượng giao dịch tài khoản ngân hàng mỗi ngày.

Biến độc lập (X):

- Ngày trong tuần (Day of the Week): Thứ trong tuần (1-7, với 1 là Chủ Nhật).
- Lãi suất (Interest Rate): Lãi suất hiện tại.
- Số lượng người sử dụng ứng dụng di động (Mobile App Users): Số lượng người sử dụng ứng dụng di động của ngân hàng.

Phương trình mô hình hồi quy Poisson có thể có dạng như sau:

$$\log(\lambda) = \beta_0 + \beta_1 \times \text{Day of the Week} + \beta_2 \times \text{Interest Rate} + \beta_3 \times \text{Mobile App Users}$$

Mô hình này có thể giúp ngân hàng hiểu rõ hơn về các yếu tố nào ảnh hưởng đến số lượng giao dịch tài khoản hàng ngày và từ đó có thể thực hiện các chiến lược quảng bá, quảng cáo, hoặc cải thiện dịch vụ để tăng cường hoạt động giao dịch tài khoản ngân hàng

Dữ liệu biến độc lập X là các dữ liệu lịch sử thu thập được từ trước. Sau khi xử lý làm sạch dữ liệu, các công cụ như R, Python... sẽ giúp chúng ta tìm ra các hệ số hồi quy gần đúng nhất với bộ dữ liệu có sẵn.

Ưu điểm

Có thể dự đoán sự kiện hiếm xảy ra

Mô hình hồi quy Poisson cho phép dự đoán sự kiện hiếm xảy ra một cách chính xác hơn so với các mô hình thống kê khác. Điều này là do mô hình Poisson được thiết kế riêng cho dữ liệu đếm sự kiện

Phù hợp với dữ liệu đếm sự kiện

Mô hình hồi quy Poisson phù hợp với dữ liệu đếm sự kiện rời rạc, chẳng hạn như số lần khách hàng mua sắm trong 1 tháng. Nó mô tả tốt mối quan hệ giữa biến phụ thuộc đếm và các biến độc lập

Xử lý tốt hiện tượng quá phân tán

Mô hình Poisson xử lý tốt hiện tượng quá phân tán của dữ liệu đếm, nghĩa là phân phối dữ liệu bị lệch phải với phương sai lớn hơn giá trị trung bình

Nhược điểm

Yêu cầu dữ liệu tính phân phối Poisson

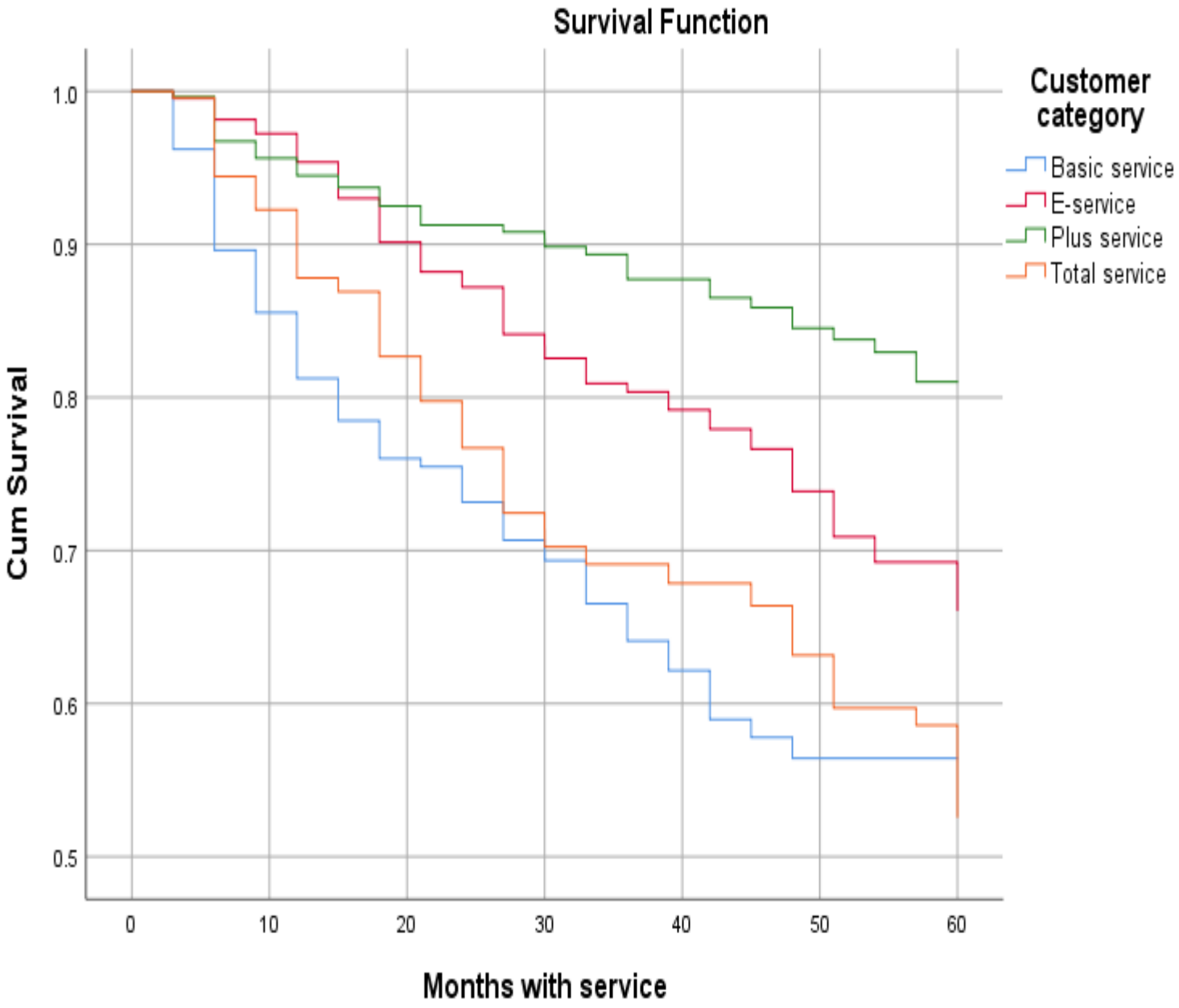
Mô hình chỉ phù hợp với dữ liệu có phân phối xác suất Poisson. Nếu dữ liệu không đúng phân phối thì mô hình sẽ không chính xác

Giả định tuyến tính

Mô hình giả định mối quan hệ tuyến tính giữa biến phụ thuộc và biến độc lập. Nếu mối quan hệ không tuyến tính thì mô hình sẽ không chính xác.

Khó giải thích các hệ số

Các hệ số trong mô hình khó có thể giải thích ý nghĩa thực tế. Điều này làm giới hạn khả năng diễn giải mô hình.



Phân tích thời gian biến cố

Survival analysis

Phân tích thời gian biến cố (phân tích sống sót) là gì?

- Định nghĩa: là một nhóm mô hình và phương pháp thống kê để dự đoán thời gian một biến cố xảy ra.
- Các yếu tố chính trong phân tích sống sót bao gồm thời gian (sau bao lâu thì biến cố xảy ra) và xác suất xảy ra biến cố trong khoảng thời gian đó.
- Một số khái niệm chính trong phân tích sống sót bao gồm:

  - Hàm số sống sót (Survival Function): Đo lường xác suất sống sót qua thời gian và làm việc đối với các đối tượng nghiên cứu $S(t) = P(T > t)$ trong đó T là thời gian đến sự kiện

  - Hàm hạn chế sống sót (Hazard Function): Đo lường tốc độ xảy ra biến cố tại một thời điểm cụ thể $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t}$

  - Kaplan-Meier Curve: Được sử dụng để ước lượng hàm số sống sót không phụ thuộc vào mô hình, thường được sử dụng để vẽ biểu đồ sống sót

  - Kiểm định log-rank: dùng để so sánh hai hoặc nhiều nhóm với nhau về thời gian đến biến cố

Phân tích sống sót phù hợp sử dụng khi yêu cầu đề bài đặt ra là phải tìm ra khoảng thời gian sau bao lâu thì một sự kiện xảy ra: ví dụ sau bao lâu sử dụng dịch vụ thì khách hàng rời bỏ, sau bao lâu chờ đợi thì khách hàng rời bỏ....

Ưu điểm

Phân tích được thời gian

Xác định được thời điểm các sự kiện quan trọng xảy ra

Xem xét nhiều yếu tố

Phân tích được các loại yếu tố khác nhau như tuổi tác, giới tính, sức khỏe ảnh hưởng đến kết quả như thế nào

Dự đoán được tương lai

Sử dụng mô hình thống kê để dự đoán khả năng sống sót trong tương lai

Nhược điểm

Yêu cầu dữ liệu tính phân phối Poisson

Mô hình chỉ phù hợp với dữ liệu có phân phối xác suất Poisson. Nếu dữ liệu không đúng phân phối thì mô hình sẽ không chính xác

Giả định tuyến tính

Mô hình giả định mối quan hệ tuyến tính giữa biến phụ thuộc và biến độc lập. Nếu mối quan hệ không tuyến tính thì mô hình sẽ không chính xác.

Khó giải thích các hệ số

Các hệ số trong mô hình khó có thể giải thích ý nghĩa thực tế. Điều này làm giới hạn khả năng diễn giải mô hình.

Tổng kết phần 2

Keywords:

Trước khi thực hiện hồi quy...

...nên vẽ biểu đồ phân tán để nhận dạng dữ liệu và chọn kiểu mô hình phù hợp

Dữ liệu nào...

... mô hình đó

Các mô hình đều có ưu/nhược điểm ...

...do đó cần sử dụng kiến thức chuyên môn để lý giải kết quả phù hợp

Nội dung

1.Tổng quan về phân tích và mô hình trong phân tích dữ liệu

2. Mô hình thường dùng cho phòng SBD

3. Tổng kết

Nội dung tóm tắt:

Tổng kết phần 1

Keywords:

Phân tích tương quan (correlation analysis) và tiên lượng (prediction)...
...thường "song hành" cùng các mô hình hồi quy

Mô hình ...
...là một phương trình để mô tả mối liên quan giữa các biến.

Mô hình hồi quy ...
...tuyến tính và phi tuyến tính

Tổng kết phần 2

Keywords:

Trước khi thực hiện hồi quy...
...nên vẽ biểu đồ phân tán để nhận dạng dữ liệu và chọn kiểu mô hình phù hợp

Dữ liệu nào...
... mô hình đó

Các mô hình đều có ưu/nhược điểm ...
...do đó cần sử dụng kiến thức chuyên môn để lý giải kết quả phù hợp

Điều quan trọng cần ghi nhớ: hệ số tương quan và mô hình tuyến tính chỉ phản ánh mối quan hệ tương quan thống kê (ví dụ A đồng biến/nghịch biến với B ở mức độ nào...) chứ không phản ánh mối quan hệ nhân quả (A suy ra B, hoặc nếu có A sẽ dẫn tới B...)

THANK YOU