BÁO CÁO 1 - NHÓM D

Contents

Thành viên:
1. Đỗ Thị Thanh Thảo (23C23009)
2. Nguyễn Kim Anh (23C23004)
3. Nguyễn Bích Trâm (23C23010)
4. Trần Thị Thuận (23C23002)

I. Giới thiệu:

1. Bài toán:

Bộ dữ liệu **"Churn_Modelling"** chứa thông tin về khách hàng và được sử dụng để phân tích hành vi khách hàng và tìm hiểu lý do khiến khách hàng rời bỏ dịch vụ.

Cấu trúc dữ liệu:

RowNumber: Chỉ mục của từng dòng dữ liệu (không ảnh hưởng đến phân tích).
CustomerId: Mã định danh của khách hàng.
Surname: Họ của khách hàng.
CreditScore: Điểm tín dụng, đánh giá khả năng tài chính của khách hàng.
Geography: Quốc gia nơi khách hàng sinh sống.
Gender: Giới tính của khách hàng.
Age: Tuổi của khách hàng.
Tenure: Thời gian khách hàng đã sử dụng dịch vụ (tính bằng năm).
Balance: Số dư tài khoản ngân hàng.
NumOfProducts: Số sản phẩm mà khách hàng sử dụng.
HasCrCard: Khách hàng có thẻ tín dụng hay không (1 = Có, 0 = Không).
IsActiveMember: Khách hàng có phải là thành viên hoạt động không (1 = Có, 0 = Không).
EstimatedSalary: Mức lương ước tính của khách hàng.
Exited: Biến mục tiêu (Target Variable):
  1: Khách hàng đã rời bỏ dịch vụ (churn).
  0: Khách hàng vẫn tiếp tục sử dụng dịch vụ.

2. Cơ sở lí thuyết:

Mô hình hồi quy đối với biến binomial có dạng:

$$log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

Binomial là một họ phân phối mũ phân tán có dạng

$$f_Z(z_i) = \exp\left\{z_i\theta_i - \log\left(1+e^{\theta_i}\right) + \log\left(C_{y_i}^{z_i m_i}\right)\right\}$$

Với:

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$$
$$b(\theta_i) = \log\left(1+e^{\theta_i}\right)$$
$$a(\phi) = 1$$
$$c(y_i,\phi) = \log\left(C_{y_i}^{z_i m_i}\right)$$

Ma trận hiệp phương sai của $\hat{\beta}$

Ma trận hiệp phương sai của $\hat{\beta}$ có công thức tổng quát $\widehat{\mathrm{Var}}(\hat{\beta}) = a(\phi_0)\left(\mathbf{X}^\top \mathbf{W}(\hat{\beta})\mathbf{X}\right)^{-1}$.

Với $a(\phi)=1$, ma trận hiệp phương sai của $\hat{\beta}$ của biến nhị thức là:

$\widehat{\mathrm{Var}}(\hat{\beta}) = \left(\mathbf{X}^\top \mathbf{W}(\hat{\beta})\mathbf{X}\right)^{-1} = \left\{X^{\mathrm{T}}\operatorname{Diag}\left[n_i\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right]X\right\}^{-1}$

Khoảng tin cậy cho hệ số cho mô hình:

Như ta đã được học ở phần trước, khoảng tin cậy của các hệ số mô hình β_j được xây dựng dựa trên tính chất tiệm cận phân phối chuẩn của ước lượng $\hat{\beta}_j$, tức là:

$$\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\phi_0 v_j}} \xrightarrow{d} \mathcal{N}(0,1),$$

Tương đương với:

$$\frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{v_j}} \xrightarrow{d} \mathcal{N}(0,1),$$

(Do với phân phối nhị thức thì ϕ_0 =1) trong đó, v_j là phương sai tiệm cận của $\hat{\beta}_j$, và được xác định bởi thành phần đường chéo thứ j của ma trận $\left(\mathbf{X}^\top \mathbf{W}(\hat{\beta})\mathbf{X}\right)^{-1}$, với

$$\mathbf{W}(\hat{\beta}) = \begin{pmatrix} W_1(\hat{\beta}) & 0 & \ldots & 0 \\ 0 & W_2(\hat{\beta}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & W_n(\hat{\beta}) \end{pmatrix} = \left\{X^{\mathrm{T}}\operatorname{Diag}\left[n_i\hat{\mu}_i\left(1-\hat{\mu}_i\right)\right]X\right\}^{-1}$$

thành phần $W_i(\hat{\beta}) = \dfrac{1}{V_i\left(\hat{\mu}_i\right)\left(g_i'\left(\hat{\mu}_i\right)\right)^2} = n_i\hat{\mu}_i\left(1-\hat{\mu}_i\right)$. Khoảng tin cậy $100\times\alpha\%$ của β_j là

$$\left(\hat{\beta}_j - z_{(1+\alpha)/2}\sqrt{v_j}, \hat{\beta}_j + z_{(1+\alpha)/2}\sqrt{v_j}\right)$$

với $z_{1-\alpha/2}$ là phân vị thứ $1-\alpha/2$ của phân phối chuẩn $\mathcal{N}(0,1)$ và v_j là thành phần đường chéo thứ j của ma trận $\left(\mathbf{X}^\top \mathbf{W}(\hat{\beta})\mathbf{X}\right)^{-1}$.

Tính chất tiệm cận của tiên đoán η

Xét tổ hợp tuyến tính $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j = x\beta$

Ước lượng của nó là $\hat{\eta} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j = x\hat{\beta}$

Ta có

$$\frac{\hat{\eta} - \eta_0}{\sqrt{\operatorname{Var}(\hat{\eta})}} \xrightarrow{d} \mathcal{N}(0,1)$$

Ước lượng ma trận phương sai của $\hat{\eta}_i$

$\widehat{\operatorname{Var}}(\hat{\eta}) = a(\phi_0)\, x \left(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X}\right)^{-1} x^\top = x \left(\mathbf{X}^\top \mathbf{W}(\hat{\beta}) \mathbf{X}\right)^{-1} x^\top = x \left\{X^{\mathrm{T}} \operatorname{Diag}\left[n_i \hat{\pi}_i \left(1 - \hat{\pi}_i\right)\right] X\right\}^{-1} x^\top$

Với khoảng tin cậy cho $\hat{\eta}_i$ là

$$\left(\hat{\eta} - z_{1-\alpha/2}\sqrt{\widehat{\operatorname{Var}}(\hat{\eta})}, \hat{\eta} + z_{1-\alpha/2}\sqrt{\widehat{\operatorname{Var}}(\hat{\eta})}\right)$$

Khoảng tin cậy cho tiên đoán trung bình

Hàm liên kết

$\eta = \log\left(\frac{\mu}{1-\mu}\right)$

Ta có:

$\eta = \log\left(\frac{\mu}{1-\mu}\right) \Rightarrow \frac{\mu}{1-\mu} = \exp\eta$

$\Rightarrow \mu = (1-\mu)\exp\eta$

$\Rightarrow \mu = \exp\eta - \mu\exp\eta$

$\Rightarrow \mu(1+\exp\eta) = \exp\eta$

$\Rightarrow \mu = \frac{\exp\eta}{1+\exp\eta}$

Vậy hàm $g = \frac{\exp\eta}{1+\exp\eta}$

$\Rightarrow$ khoảng tin cậy $100 \times (1-\alpha)\%$ cho μ được xây dựng bởi áp dụng $g^{-1}(\cdot)$ lên khoảng tin cậy của η:

$\left(g^{-1}\left(\hat{\eta}_L\right), g^{-1}\left(\hat{\eta}_U\right)\right)$

Vậy với hàm liên kết logistic, khoảng tin cậy $100 \times (1-\alpha)\%$ cho $\hat{\mu}$ là $\left(\frac{\exp\left(\hat{\eta}_L\right)}{1+\exp\left(\hat{\eta}_L\right)}, \frac{\exp\left(\hat{\eta}_U\right)}{1+\exp\left(\hat{\eta}_U\right)}\right)$