BÁO CÁO 2 - NHÓM D

Contents

I. Giới thiệu: **2**
Bài toán: . . . 2
Cấu trúc dữ liệu: . . . 2

II. Cơ sở lí thuyết: **3**
1. Ma trận hiệp phương sai của $\hat{\beta}$ . . . 3
2. Khoảng tin cậy cho hệ số cho mô hình: . . . 3
3. Tính chất tiệm cận của tiên đoán η . . . 4
4. Ước lượng ma trận phương sai của $\widehat{\eta}_i$ . . . 4
5. Khoảng tin cậy cho tiên đoán trung bình . . . 4
6. Kiểm định giả thuyết cho mô hình hồi quy Logistic . . . 5
    6.1. Kiểm định giả thuyết cho từng hệ số . . . 5
    6.2. Kiểm định giả thuyết tổng quát . . . 5

III. Bài toán thực tế **6**
Xây dựng mô hình . . . 6
Nhận xét kết quả của mô hình . . . 11

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:

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.

II. 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:

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

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

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

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

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

2. 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

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

thành phần $W_i(\hat{\beta}) = \dfrac{1}{V_i(\hat{\mu}_i)\left(g_i'(\hat{\mu}_i)\right)^2} = n_i\hat{\mu}_i\left(1-\hat{\mu}_i\right)$. 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 $\left(\mathbf{X}^\top \mathbf{W}(\hat{\beta})\mathbf{X}\right)^{-1}$.

3. 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

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

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

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

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

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

5. 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:
$(g^{-1}(\hat{\eta}_L), g^{-1}(\hat{\eta}_U))$

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

6. Kiểm định giả thuyết cho mô hình hồi quy Logistic

6.1. Kiểm định giả thuyết cho từng hệ số

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Tổng quát:

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

Do mô hình hồi quy logistic có $a(\Phi) = 1$, nếu H_0 đúng, ta có thống kê của kiểm định là:

$$Z = \frac{\hat{\beta}_j}{\sqrt{v_j}} \xrightarrow{d} N(0,1)$$

Với v_j được tính như ở phần trên.

p-value =

$$Pr(|Z| > |Z_{obs}|)$$

6.2. Kiểm định giả thuyết tổng quát

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \\ H_1 : \exists \beta_j \neq 0 \end{cases}$$

$a(\Phi) = 1$ biết trước, ta có thống kê của kiểm định:

$$W = D(Y, \hat{\mu_0}) - D(Y, \hat{\mu_1}) \sim \chi^2_p$$

Trong đó:

$$D(Y, \mu) = \sum_{i=1}^{n} d(Y_i, \mu_i)$$

Với phân phối nhị thức:

$$d(y_i, \hat{\mu_i}) = 2\{y_i log(\frac{y_i}{\hat{\mu_i}}) + (1 - y_i) log(\frac{1 - y_i}{1 - \hat{\mu_i}})\}$$

III. Bài toán thực tế

Xây dựng mô hình

```
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ggplot2)
```

Dự đoán khách hàng có rời bỏ (churn) ngân hàng hay không.

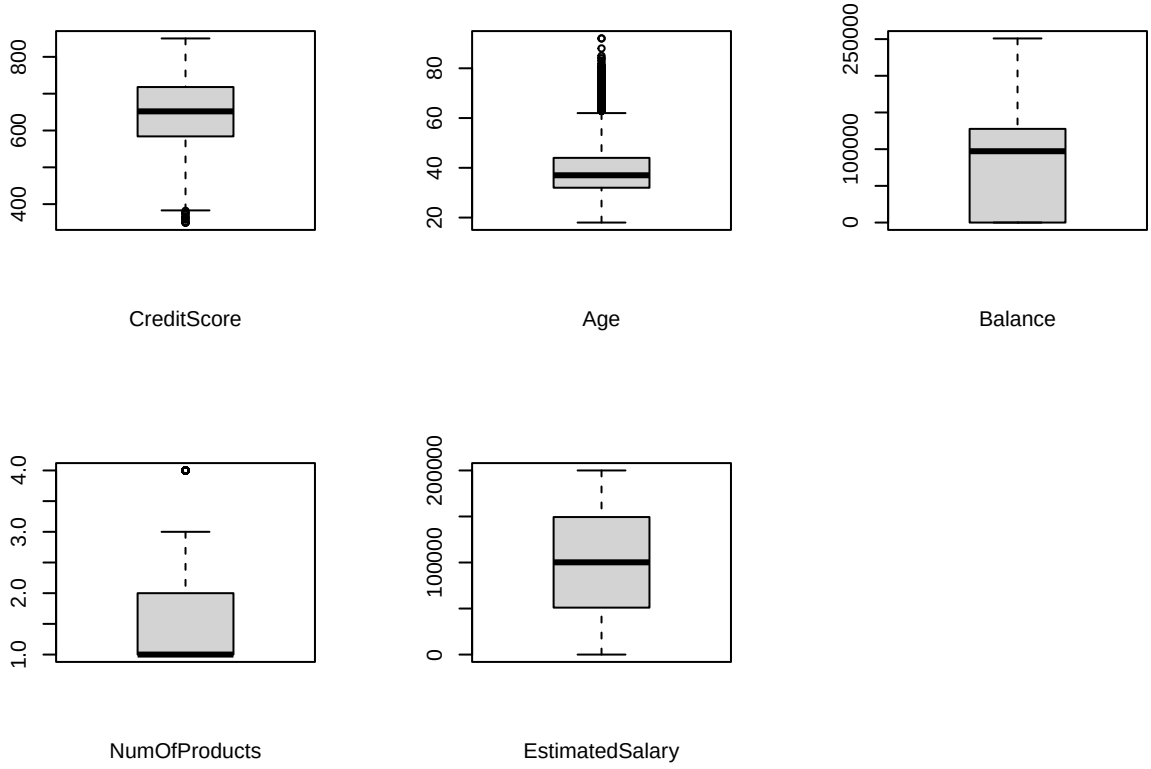
```
data <- read.csv('./Churn_Modelling.csv')
head(data)
```

```
##   RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
## 1         1   15634602 Hargrave         619    France Female  42      2
## 2         2   15647311     Hill         608     Spain Female  41      1
## 3         3   15619304     Onio         502    France Female  42      8
## 4         4   15701354     Boni         699    France Female  39      1
## 5         5   15737888 Mitchell         850     Spain Female  43      2
## 6         6   15574012      Chu         645     Spain   Male  44      8
##     Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## 1      0.00             1         1              1       101348.88      1
## 2  83807.86             1         0              1       112542.58      0
## 3 159660.80             3         1              0       113931.57      1
## 4      0.00             2         0              0        93826.63      0
## 5 125510.82             1        NA              1        79084.10      0
## 6 113755.78             2         1              0       149756.71      1
```

```
# Loại bỏ những cột không có ý nghĩa trong mô hình
# 'CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'Balance', 'NumOfProducts','EstimatedSalary','Exi
data <- subset(data, select=-c(RowNumber, CustomerId, Surname))
head(data)
```

```
##   CreditScore Geography Gender Age Tenure   Balance NumOfProducts HasCrCard
## 1         619    France Female  42      2      0.00             1         1
## 2         608     Spain Female  41      1  83807.86             1         0
## 3         502    France Female  42      8 159660.80             3         1
## 4         699    France Female  39      1      0.00             2         0
## 5         850     Spain Female  43      2 125510.82             1        NA
## 6         645     Spain   Male  44      8 113755.78             2         1
##   IsActiveMember EstimatedSalary Exited
## 1              1       101348.88      1
## 2              1       112542.58      0
## 3              0       113931.57      1
## 4              0        93826.63      0
## 5              1        79084.10      0
## 6              0       149756.71      1
```

```
numeric_cols = c('CreditScore', 'Age', 'Balance', 'NumOfProducts', 'EstimatedSalary')
par(mfrow=c(2, 3))
for (i in numeric_cols){
  boxplot(data[[i]], xlab=i)
}
```



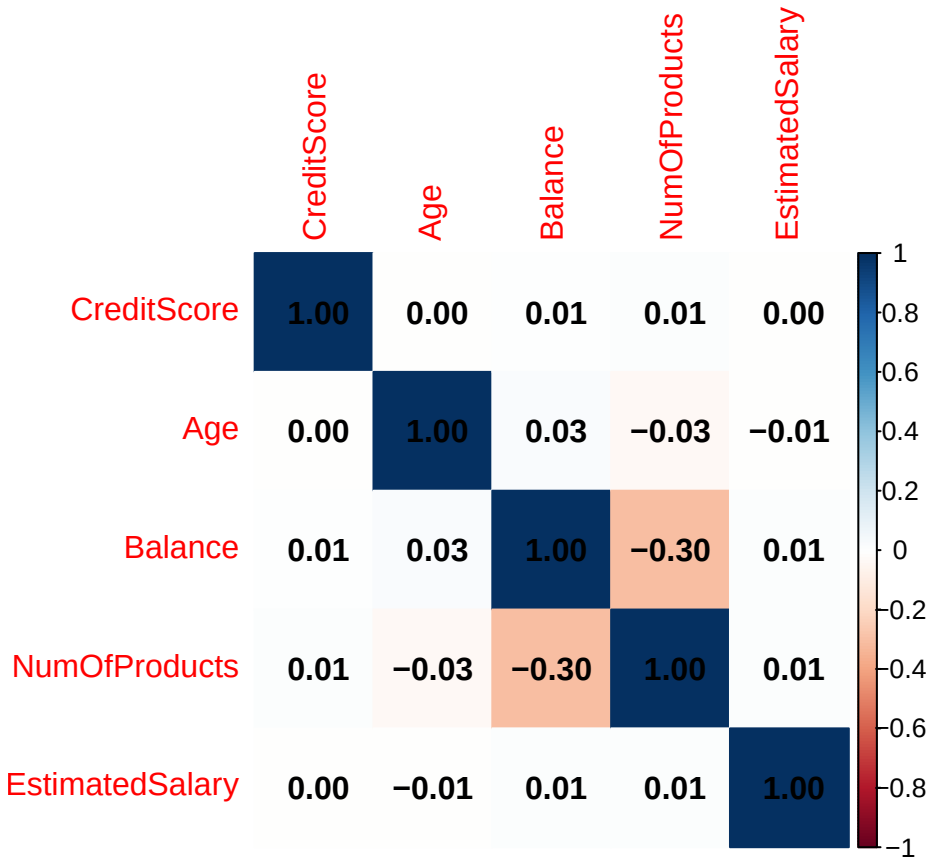
Không có dữ liệu outlier trong dữ liệu

```
colSums(is.na(data))/nrow(data)
```

```
##      CreditScore        Geography           Gender              Age 
##        0.000e+00        0.000e+00        0.000e+00        9.998e-05 
##           Tenure          Balance    NumOfProducts        HasCrCard 
##        0.000e+00        0.000e+00        0.000e+00        9.998e-05 
##   IsActiveMember  EstimatedSalary           Exited 
##        9.998e-05        0.000e+00        0.000e+00
```

```
# loại null
data <- subset(data, is.na(data$Age) == FALSE)
```

```
# kiểm tra sự tương quan
numeric_cols = c('CreditScore', 'Age', 'Balance', 'NumOfProducts', 'EstimatedSalary')
corrplot(cor(data[,numeric_cols]), addCoef.col = 'black', method="color")
```



Không có sự tương quan mạnh giữa các biến

```
# Xử lý tạo biến giả cho các biến định tính
data$Geography = relevel(as.factor(data$Geography), ref=1)
data$Gender = relevel(as.factor(data$Gender), ref=1)
data$Tenure = relevel(as.factor(data$Tenure), ref=1)
data$HasCrCard = relevel(as.factor(data$HasCrCard), ref=1)
data$IsActiveMember = relevel(as.factor(data$IsActiveMember), ref=1)
```

```
logit_model <- glm(Exited ~., family="binomial", data=data)
summary(logit_model)
```

```
##
## Call:
## glm(formula = Exited ~ ., family = "binomial", data = data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.251e+01  1.970e+02  -0.063   0.9494
## CreditScore      -6.603e-04  2.806e-04  -2.354   0.0186 *
## GeographyFrance   9.185e+00  1.970e+02   0.047   0.9628
## GeographyGermany  9.959e+00  1.970e+02   0.051   0.9597
## GeographySpain    9.222e+00  1.970e+02   0.047   0.9627
## GenderMale       -5.276e-01  5.454e-02  -9.673  < 2e-16 ***
## Age               7.270e-02  2.579e-03  28.189  < 2e-16 ***
## Tenure1          -5.519e-02  1.513e-01  -0.365   0.7153
## Tenure2          -2.397e-01  1.529e-01  -1.567   0.1171
## Tenure3          -1.197e-01  1.524e-01  -0.785   0.4325
## Tenure4          -7.553e-02  1.532e-01  -0.493   0.6220
## Tenure5          -1.903e-01  1.530e-01  -1.244   0.2137
```

```
## Tenure6          -1.474e-01  1.540e-01  -0.957   0.3387
## Tenure7          -3.175e-01  1.553e-01  -2.044   0.0409 *
## Tenure8          -2.255e-01  1.538e-01  -1.466   0.1425
## Tenure9          -1.699e-01  1.530e-01  -1.110   0.2668
## Tenure10         -2.076e-01  1.764e-01  -1.177   0.2392
## Balance           2.655e-06  5.147e-07   5.157 2.51e-07 ***
## NumOfProducts    -9.610e-02  4.722e-02  -2.035   0.0419 *
## HasCrCard1       -4.382e-02  5.939e-02  -0.738   0.4607
## IsActiveMember1  -1.075e+00  5.773e-02 -18.616  < 2e-16 ***
## EstimatedSalary   4.792e-07  4.739e-07   1.011   0.3119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10112.1  on 9998  degrees of freedom
## Residual deviance:  8556.7  on 9977  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 8600.7
##
## Number of Fisher Scoring iterations: 10
```

Nhận xét: Với ngưỡng anpha = 0.05, chỉ có Credit score, GenderMale, Age, Tenure7, Balance, NumOfProducts và IsActiveMember1 là có ý nghĩa trong mô hình

```r
confint(logit_model)
```

```
## Waiting for profiling to be done...
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
##                          2.5 %        97.5 %
## (Intercept)                 NA  2.762040e+01
## CreditScore      -1.210467e-03 -1.105821e-04
## GeographyFrance  -3.095710e+01            NA
## GeographyGermany -3.018185e+01            NA
## GeographySpain   -3.091985e+01            NA
## GenderMale       -6.346203e-01 -4.208038e-01
## Age               6.766649e-02  7.777703e-02
## Tenure1          -3.494935e-01  2.440624e-01
## Tenure2          -5.374235e-01  6.253232e-02
## Tenure3          -4.163091e-01  1.816845e-01
## Tenure4          -3.737352e-01  2.272787e-01
## Tenure5          -4.880522e-01  1.120965e-01
## Tenure6          -4.472801e-01  1.569434e-01
## Tenure7          -6.200490e-01 -1.085735e-02
## Tenure8          -5.248622e-01  7.829629e-02
## Tenure9          -4.677922e-01  1.325107e-01
## Tenure10         -5.535202e-01  1.385300e-01
## Balance           1.645497e-06  3.663475e-06
## NumOfProducts    -1.889766e-01 -3.828365e-03
## HasCrCard1       -1.598464e-01  7.299393e-02
## IsActiveMember1  -1.188350e+00 -9.620336e-01
## EstimatedSalary  -4.495796e-07  1.408397e-06
```

```
W_obs <- logit_model$null.deviance - logit_model$deviance
W_obs
```

```
## [1] 1555.342
```

```
print(1 - pchisq(W_obs, df=10))
```

```
## [1] 0
```

- Mô hình:

$$log(\frac{\mu}{1-\mu}) = -12.51 - 0.0006603 * CreditScore - 0.5276 * GenderMale + 0.06766649 * Age - 0.3175 * Tenure7 + 2.655 * 10^{-6} * Balance -$$

Ý nghĩa hệ số coefficient:
- Đối với biến liên tục: Khi biến X_i tăng/giảm 1 đơn vị thì $\frac{\mu}{1-\mu}$ (odds) tăng/giảm e^{β_i} lần.
- Đối với biến rời rạc: Tỷ lệ odds của nhóm đối chiếu là cao hơn là e^{β_i} nhóm tham chiếu. Hay xác suất của nhóm đối chiếu cao hơn nhóm tham chiếu là 1-odds lần.

Nhận xét kết quả của mô hình

- Khi CreditScore tăng 1 đơn vị thì odds Exit giảm $e^{0.0006603} = 1.000661$ lần
- Khi Age tăng 1 đơn vị thì odds Exit tăng $e^{0.06766649} = 1.070008$ lần
- Khi Balance tăng 1 đơn vị thì odds Exit tăng $e^{2.655e-06} = 1.000003$ lần
- Khi số lượng NumOfProducts tăng 1 đơn vị thì odds Exit giảm $e^{0.0961} = 1.100869$ lần
- Xác suất exit của nhóm male ít hơn nhóm femal là $1 - e^{0.5276} = 69.5\%$
- Xác suất exit của nhóm Tenure7 ít hơn nhóm không phải Tenure7 là $1 - e^{0.3175} = 37.37\%$
- Xác suất exit của nhóm IsActiveMember1 ít hơn nhóm không phải IsActiveMember1 là $1 - e^{1.075} = 193\%$