

Transformació & Selecció

Wenjia Ye

2025-10-11

Transformació

S'ha decidit transformar les variables `ComplaintsCount` i `NumOfProducts` en variables categòriques, ja que en ambdós casos presenten una distribució molt desequilibrada.

En el cas de `ComplaintsCount`, la gran majoria de clients no han presentat cap queixa, mentre que només una petita proporció n'ha fet una o més. Per tant, es transforma en una variable binària amb dos nivells:

- 0: el client no s'ha queixat
- 1: el client sí que s'ha queixat

De manera similar, la variable `NumOfProducts` mostra que la major part dels clients tenen un sol producte, uns quants en tenen dos, i només una minoria en té tres o més. Per això, s'ha recodificat en tres categories:

- Clients amb 1 producte
- Clients amb 2 productes
- Clients amb 3 o més productes

En tots dos casos, la transformació de variables numèriques a categòriques permet reduir la influència de valors poc freqüents, evitar soroll estadístic i facilitar la interpretació dels resultats. A més, aquesta simplificació pot contribuir a millorar l'estabilitat i rendiment dels models predictius, especialment en aquells que treballen millor amb variables discretes o amb distribucions equilibrades.

```
load("C:/Users/34688/OneDrive - Universitat de Barcelona/Escritorio/Mineria/DATA/data_imputado.RData")
data<-data_imputado
data$ComplaintsCount_bin <- ifelse(data$ComplaintsCount == 0, 0, 1)
data$ComplaintsCount_bin <- factor(data$ComplaintsCount_bin,
                                   levels = c(0, 1),
                                   labels = c("No_queja", "Queja"))

data$NumOfProducts_grupo <- ifelse(data$NumOfProducts == 1, "1",
                                   ifelse(data$NumOfProducts == 2, "2", "3 o más"))
data$NumOfProducts_grupo <- factor(data$NumOfProducts_grupo,
                                   levels = c("1", "2", "3 o más"))
```

Selecció

Un cop imputades i analitzades les dades, tal com s'ha mencionat anteriorment, es decideix reduir el dataset centrant-se en les variables que presenten una forta relació amb la variable resposta. Tot i que anteriorment ja s'havien realitzat proves per identificar quines variables eren estadísticament significatives, les transformacions i imputacions recents poden haver modificat les distribucions i relacions originals, tot i que no ha de passar.

Per aquest motiu, es tornen a aplicar els tests estadístics corresponents:

- Test de mediana per a les variables numèriques
- Test de Chi-quadrat per a les variables categòriques

Aquesta nova avaluació garanteix que la selecció de variables sigui coherent amb el dataset final, després de totes les modificacions realitzades.

```
varCat<-c("Geography", "Gender", "MaritalStatus", "EducationLevel", "HasCrCard",
          "SavingsAccountFlag", "LoanStatus", "CustomerSegment", "Exited", "IsActiveMember", "ComplaintsClosed")
varNum<-c("Age", "CreditScore", "Tenure", "EstimatedSalary", "Balance", "TransactionFrequency", "AvgTransactionAmount")
v<-"Exited"
for (varc1 in varCat) {
  if (varc1 != v) {
    tab <- table(data[[v]], data[[varc1]]) # filas = Exited, columnas = categorías
    test <- chisq.test(tab, correct = FALSE)

    cat("\nVariable:", varc1, "\n")
    cat("Chi-squared =", round(test$statistic, 3),
        "df =", test$parameter,
        "p-value =", signif(test$p.value, 5), "\n")

    if (test$p.value < 0.05) cat("-> Diferencias significativas entre columnas\n")
  }
}
```

```
##
## Variable: Geography
## Chi-squared = 97.683 df = 2 p-value = 6.1421e-22
## -> Diferencias significativas entre columnas
##
## Variable: Gender
## Chi-squared = 55.63 df = 1 p-value = 8.746e-14
## -> Diferencias significativas entre columnas
##
## Variable: MaritalStatus
## Chi-squared = 1.883 df = 3 p-value = 0.59695
##
## Variable: EducationLevel
## Chi-squared = 3.899 df = 3 p-value = 0.27254
##
## Variable: HasCrCard
## Chi-squared = 0.425 df = 1 p-value = 0.5143
##
## Variable: SavingsAccountFlag
## Chi-squared = 0.404 df = 1 p-value = 0.52506
##
## Variable: LoanStatus
## Chi-squared = 0.218 df = 2 p-value = 0.89689
##
## Variable: CustomerSegment
## Chi-squared = 1.199 df = 2 p-value = 0.54901
##
## Variable: IsActiveMember
## Chi-squared = 74.063 df = 1 p-value = 7.5648e-18
```

```
## -> Diferencias significativas entre columnas
##
## Variable: ComplaintsCount_bin
## Chi-squared = 0.006 df = 1 p-value = 0.93943
##
## Variable: NumOfProducts_grupo
## Chi-squared = 541.239 df = 2 p-value = 2.961e-118
## -> Diferencias significativas entre columnas
```

```
resultados_mediana <- data.frame(
  Variable = character(),
  Mediana_Exited0 = numeric(),
  Mediana_Exited1 = numeric(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)

for (var in varNum) {
  medianas <- tapply(data[[var]], data$Exited, median, na.rm = TRUE)
  p_val <- wilcox.test(data[[var]] ~ data$Exited)$p.value

  resultados_mediana <- rbind(resultados_mediana, data.frame(
    Variable = var,
    Mediana_Exited0 = medianas["0"],
    Mediana_Exited1 = medianas["1"],
    p_value = p_val
  ))
}
resultados_mediana
```

	Variable	Mediana_Exited0	Mediana_Exited1	p_value
## 0	Age	36.00000	43.00000	5.036801e-82
## 01	CreditScore	652.00000	647.00000	1.239849e-01
## 02	Tenure	5.00000	5.00000	8.658442e-01
## 03	EstimatedSalary	98857.51500	103537.00500	1.401894e-01
## 04	Balance	94243.22000	107850.82000	2.264889e-13
## 05	TransactionFrequency	30.00000	30.00000	7.811747e-01
## 06	AvgTransactionAmount	99.32957	97.59089	6.395712e-01
## 07	DigitalEngagementScore	60.00000	60.00000	3.507634e-01
## 08	NetPromoterScore	8.00000	8.00000	7.625010e-01

Un cop efectuats els tests estadístics pertinents, s'ha constatat un canvi en relació amb els resultats obtinguts abans de la imputació. Les variables CreditScore i EstimateSalary han deixat de ser estadísticament significatives. Aquest resultat no és inesperat, ja que anteriorment presentaven p-valors només lleugerament inferiors a 0,05, fet que ja indicava una significació marginal.

```
data_transformada <- subset(data, select = -c(ComplaintsCount, NumOfProducts))
data_reducida <- data[, c("Geography", "Gender", "Exited", "IsActiveMember", "NumOfProducts_grupo", "Age", "Ba
```