**SESIÓN 4: PROFILING**

Una de las mayores preocupaciones en todas las empresas es evitar la fuga de clientes, este es el caso del sector bancario. Para poder tener un conocimiento fiable y poder adoptar políticas de prevención, se han listado las bajas ocurridas en un periodo de tiempo de terminado. Para ellas se ha recogido la información que se dispone del ex cliente: sus características y posición en el año anterior a la baja (posición considerada de base) y el cambio de posición ocurrido hasta los tres meses antes de la baja (diferencia entre la posición tres meses antes de la baja y la posición de base). La información se completa con una muestra aleatoria de clientes que no han sido baja en el periodo considerado, para los que se recoge la misma información.

Con los datos recogidos se ha formado el fichero “churn.txt”, conteniendo las siguientes variables:

> names(churn)

[1] "Baja" "edatcat" "sexo"

[4] "antig" "Nomina" "Pension"

[7] "Debito\_normal" "Debito\_aff" "VISA"

[10] "VISA\_aff" "MCard" "Amex"

[13] "Total\_activo" "Total\_Plazo" "Total\_Inversion"

[16] "Total\_Seguros" "Total\_Vista" "dif\_resid"

[19] "oper\_caj\_Libreta" "oper\_ven\_Libreta" "dif\_CC"

[22] "dif\_Libreta" "dif\_Plazo" "dif\_Ahorro"

[25] "dif\_Largo\_plazo" "dif\_Fondos\_inv" "dif\_Seguros"

[28] "dif\_Planes\_pension" "dif\_Hipoteca" "dif\_Prest\_personales"

1. Lea este fichero y efectúe un “summary” de los datos. ¿Detecta algún error o inconsistencia?. Si es así, corríjalo.
2. Especifique cuál es la variable de respuesta y cuáles son las explicativas y el tipo de todas ellas.
3. Efectúe una gráfica de los datos: un diagrama de barras para las variables categóricas y un histograma para las variables continuas.
4. Efectúe el “profiling” de las bajas (con la función *catdes* de la librería “FactoMineR”). Interprete el resultado.
5. Represente visualmente la relación de las variables explicativas con la variable de respuesta; para ello discretize las variables continuas (esto es, recodifíquelas según un cierto número de intervalos; tenga en cuenta el significado especial del valor 0 a la hora de establecer los intervalos de recodificación) y represente mediante barplots el *porcentaje de baja* de las modalidades de las variables categóricas (tanto las categóricas originales como las continuas recodificadas).
6. Suponga que quiere analizar la compra de un producto a partir del barrio de residencia (en clase alta o baja, indicador del poder adquisitivo del cliente). En un primer análisis se obtiene la siguiente tabla:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Compra SI | Compra NO | Total |
| Clase alta | 20 | 373 | 393 |
| Clase baja | 6 | 316 | 322 |

En su opinión, ¿el poder adquisitivo del cliente, tiene alguna influencia sobre la compra o no del producto? (Responda sólo calculando las probabilidades, sin realizar la prueba de hipótesis de igualdad entre ambas probabilidades).

Un empleado senior de la compañía nos sugiere profundizar más en el análisis y tener en cuenta la edad de los clientes. Cruzando por edad (adulto o joven) los dos tipos de barrio mencionados, obtenemos las siguientes tablas:

|  |  |  |  |
| --- | --- | --- | --- |
| ADULTOS | Compra SI | Compra NO | Total |
| Clase alta | 3 | 176 | 179 |
| Clase baja | 4 | 293 | 297 |
| JOVENES | Compra SI | Compra NO | Total |
| Clase alta | 17 | 197 | 214 |
| Clase baja | 2 | 23 | 25 |

¿Tenía razón el empleado de que era conveniente tener en cuenta la edad?. ¿Cuál de los dos factores, el barrio de residencia o la edad, es el determinante en la compra del producto en cuestión?

*Utilice R para la solución de los ejercicios*

*Para la recodificación de las variables continuas, puede utilizar el siguiente código R:*

churn2 = churn

# recoding of continuous variables

churn2$Rec\_tot\_activo = cut(churn$Total\_activo, breaks=c(0,0.0001,150,400,1000,3000,99000),include.lowest=T)

churn2$Rec\_tot\_plazo = cut(churn$Total\_Plazo, breaks=c(0,0.0001,700,2000,4000,8000,99000),include.lowest=T)

churn2$Rec\_tot\_inversion = cut(churn$Total\_Inversion, breaks=c(0,0.0001,700,2000,4000,8000,99000),include.lowest=T)

churn2$Rec\_tot\_seguros = cut(churn$Total\_Seguros, breaks=c(0,0.0001,150,400,1000,3000,99000),include.lowest=T)

churn2$Rec\_tot\_vista = cut(churn$Total\_Vista, breaks=c(0,0.0001,50,150,400,1000,99000),include.lowest=T)

churn2$Rec\_oper\_caj\_Libreta = cut(churn$oper\_caj\_Libreta, breaks=c(-9000,-100,-20,-0.0001,0,20,100,9000))

churn2$Rec\_oper\_ven\_Libreta = cut(churn$oper\_ven\_Libreta, breaks=c(-9000,-100,-20,-0.0001,0,20,100,9000))

churn2$Rec\_dif\_CC= cut(churn$dif\_CC, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Libreta= cut(churn$dif\_Libreta, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Plazo= cut(churn$dif\_Plazo, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Ahorro= cut(churn$dif\_Ahorro, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Largo\_plazo= cut(churn$dif\_Largo\_plazo, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Fondos\_inv= cut(churn$dif\_Fondos\_inv, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Seguros= cut(churn$dif\_Seguros, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Planes\_pension= cut(churn$dif\_Planes\_pension, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Hipoteca= cut(churn$dif\_Hipoteca, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))

churn2$Rec\_dif\_Prest\_personales= cut(churn$dif\_Prest\_personales, breaks=c(-99000,-100,-0.0001,0,20,200,1000,99000))