



Facultad de Ciencias

UNAM

ESTADÍSTICA 3. MODELOS DE SUPERVIVENCIA Y SERIES DE TIEMPO

Proyecto final

ANÁLISIS DE DATOS DE CLIENTES BANCARIOS

Enríquez Hernández Leobardo
Tlahuiz Tenorio Saúl Giovanni

7 de mayo de 2024

Índice

Introducción.	1
1. Estadística descriptiva y procesamiento de datos.	1
Selección de variables y verificación de datos faltantes.	1
Resumen de los datos numéricos y detección de outliers.	2
Variables categóricas.	3
2. Relación entre variable de censura y covariables.	5
3. El problema de supervivencia.	6
Conclusiones	6
Referencias	7

Introducción.

Este documento tiene dos principales objetivos generales, mostrar algunos elementos estadísticos del análisis de supervivencia y al mismo tiempo con base en unos datos de clientes bancarios hacer algunas predicciones sobre la pérdida de clientes.

En la primera sección se hace la estadística descriptiva de los datos con los que se trabajará, para dar un contexto y un panorama general de la naturaleza y características de las variables. En la sección dos, haremos procesamiento de los datos en caso de que sea necesario por ejemplo tratar con valores perdidos, valores atípicos, etc. En la tercera sección se plantea y desarrolla el problema de supervivencia. Y finalmente, en la cuarta sección se presentan las principales conclusiones del trabajo.

1. Estadística descriptiva y procesamiento de datos.

La base de datos es de clientes de un banco con las siguientes variables:

- id: número de fila de la observación, comenzando por el 0.
- CustomerId: número de cuenta del cliente.
- Surname: apellido.
- CreditScore: puntaje de crédito.
- Geography: país de residencia.
- Gender: género del cliente.
- Age: edad del cliente.
- Tenure: cuántos años ha tenido cuenta bancaria en el Banco.
- Balance: saldo de la cuenta.
- NumOfProducts: número de productos bancarios en el Banco.
- HasCrCard: si tiene o no tarjeta de crédito (sí=1).
- IsActiveMember: si es miembro activo del banco (sí=1).
- EstimatedSalary: salario estimado.
- Exited: si el cliente ha dejado el banco por algún periodo (sí=1).

Selección de variables y verificación de datos faltantes.

Primero tomaremos un subconjunto del conjunto total de variables, omitiremos variables que no utilizaremos en el análisis tales como id, CustomerId, y Surname. Luego mostraremos en el siguiente cuadro que no hay datos faltantes (NA's) para las variables elegidas.

CreditScore : 0	Geography: 0	Gender: 0	Age: 0
Tenure : 0	Balance: 0	NumOfProducts: 0	HasCrCard: 0
IsActiveMember : 0	EstimatedSalary: 0	Exited: 0	

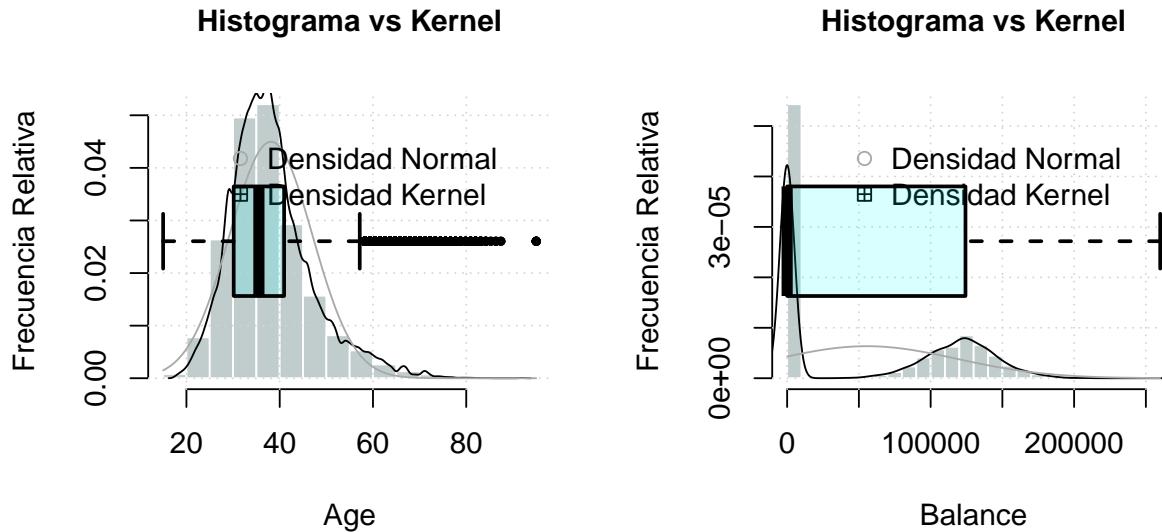
Resumen de los datos numéricos y detección de outliers.

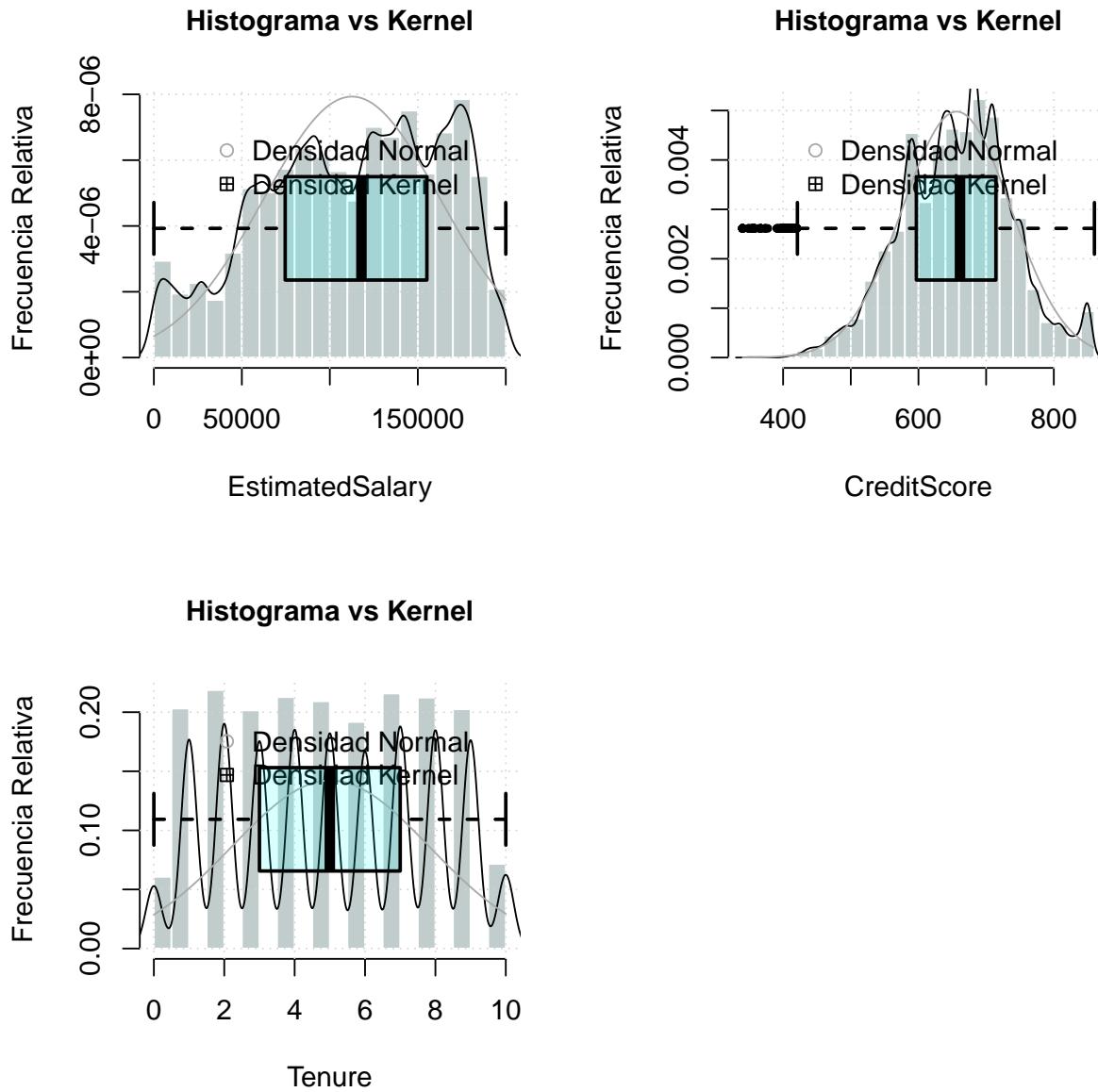
A continuación se muestra la estadística descriptiva de los valores numéricos relevantes. Son 165,034 observaciones, con edades entre 18 y 92 años, con un balance de 0 a 250,898 unidades monetarias, con salario estimado de entre 11,58 a 199,992,5, un score de crédito de 350 a 850, y tenencia de cuenta bancaria de 0 a 10 años. El promedio de edad es de 38 años, con un balance promedio de 55,478 unidades monetarias, un promedio de salario estimado de 112,575,8, un promedio de credit score de 656,45, y un promedio de tenencia de cuenta de 5 años.

Statistic	N	Mean	St. Dev.	Min	Max
Age	165,034	38.126	8.867	18.000	92.000
Balance	165,034	55,478.090	62,817.660	0.000	250,898.100
EstimatedSalary	165,034	112,574.800	50,292.870	11.580	199,992.500
CreditScore	165,034	656.454	80.103	350	850
Tenure	165,034	5.020	2.806	0	10

A continuación se muestran los histogramas de frecuencias relativas, distribuciones o densidades y boxplot conjuntas para las variables numéricas. Se pueden observar distribuciones multimodales, a excepción de la variable de edad. Por ejemplo, el balance tiene un sesgo muy notable a la izquierda, mientras que el salario estimado presenta una distribución multimodal muy irregular, y el credit score por ser una variable discreta, se percibe con varios saltos.

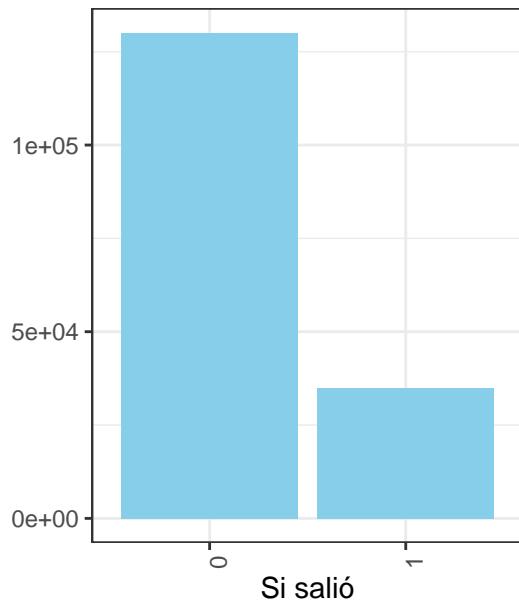
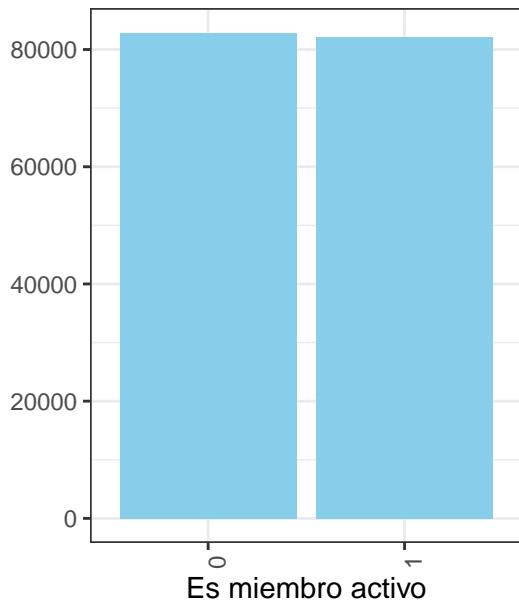
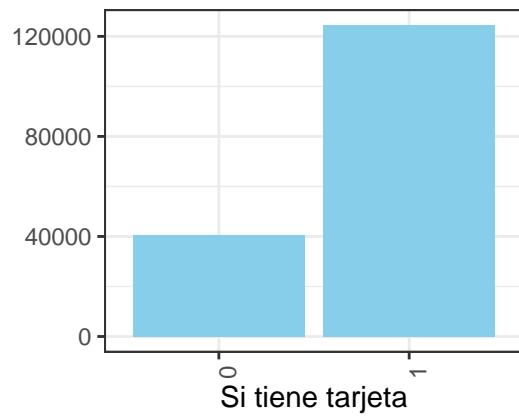
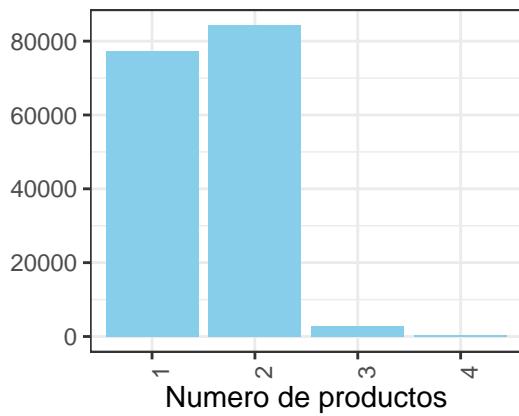
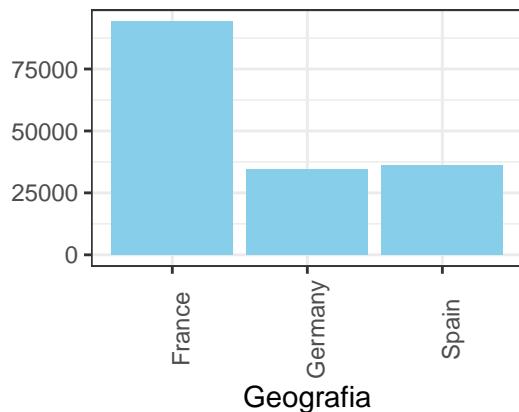
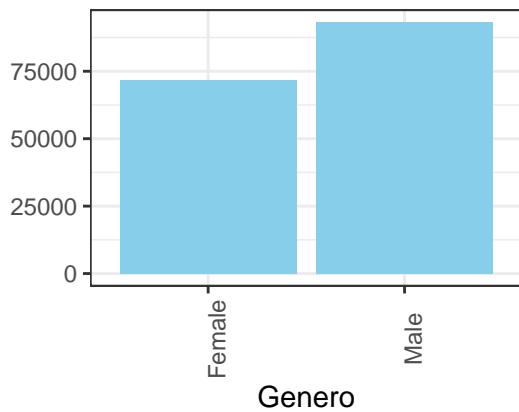
No se observan outliers importantes, en el caso de la edad, hay valores de 92 años, pero no los consideramos outliers, por lo que no quitamos estos valores que corresponde a 11 observaciones.





Variables categóricas.

A continuación presentaremos los datos categóricos, por su frecuencia absoluta en la base de datos. En la base se tienen 71,884 mujeres y 93,150 hombres; 94,215 son de Francia, 34,606 de Alemania y 36,213 de España; 77,374 tiene un solo producto bancario, 84,291 tienen dos, mientras que 2,894 tienen tres productos, y 475 tienen 4 productos; 40,606 no tiene tarjeta de crédito y 124,428 sí tiene; 82,885 no es miembro activo y 82,149 son miembros activos; y 130,113 permanecen con el banco, mientras que 34,921 salieron del banco por algún periodo.



2. Relación entre variable de censura y covariables.

```

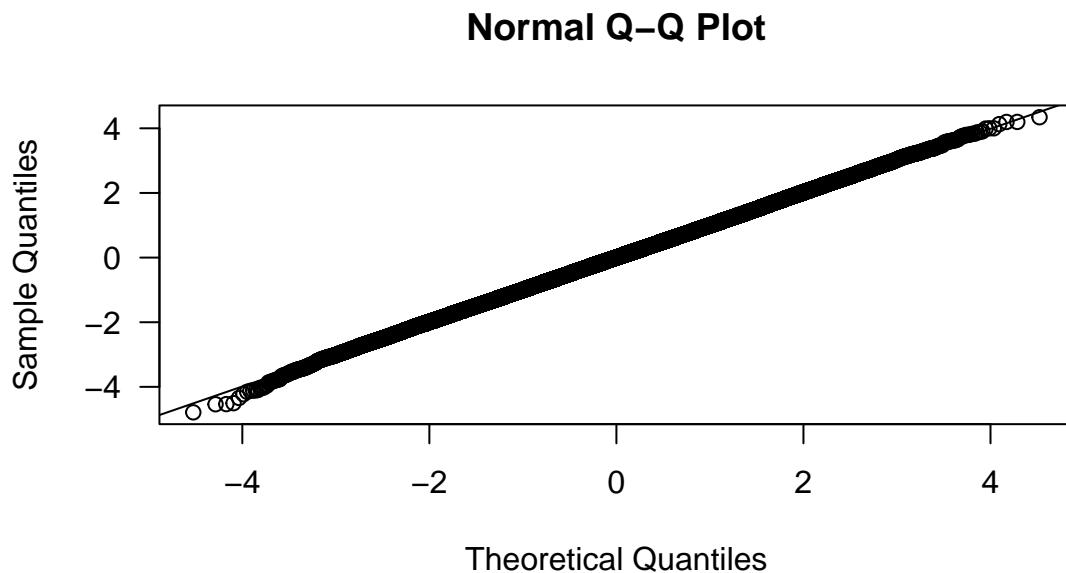
## 
## Call:
## glm(formula = Exited ~ ., family = binomial(link = "logit"),
##      data = datos)
## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.048e+00 8.553e-02 -23.941 < 2e-16 ***
## CreditScore          -7.665e-04 9.347e-05  -8.201 2.39e-16 ***
## GeographyGermany    1.279e+00 2.077e-02   61.572 < 2e-16 ***
## GeographySpain       2.512e-02 1.986e-02    1.265   0.206
## GenderMale           -6.655e-01 1.509e-02 -44.110 < 2e-16 ***
## Age                  8.982e-02 8.409e-04 106.817 < 2e-16 ***
## Tenure               -1.602e-02 2.670e-03  -6.000 1.97e-09 ***
## Balance              -5.312e-06 1.505e-07 -35.303 < 2e-16 ***
## NumOfProducts.L      2.658e+00 1.032e-01   25.746 < 2e-16 ***
## NumOfProducts.Q      1.001e+00 8.280e-02   12.089 < 2e-16 ***
## NumOfProducts.C      -2.617e+00 5.518e-02 -47.433 < 2e-16 ***
## HasCrCard1           -1.665e-01 1.721e-02  -9.676 < 2e-16 ***
## IsActiveMember1     -1.261e+00 1.600e-02 -78.849 < 2e-16 ***
## EstimatedSalary      9.914e-07 1.499e-07    6.613 3.77e-11 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 170337  on 165033  degrees of freedom
## Residual deviance: 113578  on 165020  degrees of freedom
## AIC: 113606
## 
## Number of Fisher Scoring iterations: 6
## 
## Call:
## glm(formula = Exited ~ ., family = binomial(link = "probit"),
##      data = datos)
## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.161e+00 4.715e-02 -24.612 < 2e-16 ***
## CreditScore          -4.097e-04 5.191e-05  -7.894 2.94e-15 ***
## GeographyGermany    7.093e-01 1.166e-02   60.842 < 2e-16 ***
## GeographySpain       1.583e-02 1.090e-02    1.453   0.146
## GenderMale           -3.718e-01 8.361e-03 -44.467 < 2e-16 ***
## Age                  4.972e-02 4.633e-04 107.336 < 2e-16 ***
## Tenure               -8.908e-03 1.483e-03  -6.008 1.88e-09 ***
## Balance              -2.664e-06 8.398e-08 -31.725 < 2e-16 ***
## NumOfProducts.L      1.485e+00 5.494e-02   27.024 < 2e-16 ***
## NumOfProducts.Q      5.303e-01 4.407e-02   12.034 < 2e-16 ***
## NumOfProducts.C      -1.448e+00 2.927e-02 -49.479 < 2e-16 ***
## HasCrCard1           -9.929e-02 9.564e-03 -10.381 < 2e-16 ***
## IsActiveMember1     -6.876e-01 8.688e-03 -79.145 < 2e-16 ***
## EstimatedSalary      5.142e-07 8.319e-08    6.181 6.35e-10 ***

```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 170337 on 165033 degrees of freedom
## Residual deviance: 114208 on 165020 degrees of freedom
## AIC: 114236
##
## Number of Fisher Scoring iterations: 6

```



En la prueba de normalidad Lilliefors (Kolmogorov-Smirnov) **normality test** tenemos que el p-value es de 0.0397619708850528, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, para la prueba de normalidad de Shapiro-Wilk **normality test** el p-value es de 0.407426456609919, lo que también no rechaza la hipótesis nula de normalidad. Esto se observa en la siguiente Gráfica.

3. El problema de supervivencia.

Conclusiones

Referencias