Actividad_2-Grupo3-Lote3

Leticia Florido, Iván Gómez, Alejandro Cita, Alejandro Campos

2026-01-05

Introducción

Para esta actividad, utilizaremos un enfoque de Ciencia de Datos Mínimo Viable. Nuestro objetivo es predecir el abandono de clientes (Churn) utilizando Regresión Logística y Árboles de Decisión.

```
## [1] "/Users/alejandrocamposlamas/Library/CloudStorage/GoogleDrive-alejandro.camposla@gmail.com/My Dr
```

2. Carga y breve exploración para primera limpieza de Datos

Cargamos los datos tratando de forzar las columnas al tipo adecuado e inspeccionamos el tipo de dato que tiene cada columna. Además seleccionaremos solamente las columnas solicitadas en la actividad para el modelado.

```
## 'data.frame':    7032 obs. of  23 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID_cliente           : Factor w/ 7032 levels "0002-ORFBO","0003-MKNFE",..: 5366 3954 2559 5525 65
##  $ Sexo                 : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ Jubilado             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Socio                : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Empleado             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ Meses_alta           : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ Servicio_telefonico  : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ Lineas_multiples     : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
##  $ Servicio_Internet    : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ Seguridad_Online     : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
##  $ CopiaSeguridad_Online : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
##  $ Proteccion_dispositivo: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
##  $ Soporte_tecnico      : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ Television_carta     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
##  $ Peliculas_carta      : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
##  $ Contrato             : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ Factura_digital      : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ Metodo_pago          : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ..
##  $ Gasto_mensual        : num  29.9 57 53.9 42.3 70.7 ...
##  $ Gasto_total          : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Abandono             : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
##  $ meses_alta_cut       : Factor w/ 6 levels "<1","1-6","18-42",..: 1 3 2 4 2 5 3 5 3 6 ...
```

```
## Rows: 7,032
## Columns: 8
## $ Abandono             <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, ~
## $ Contrato             <fct> Month-to-month, One year, Month-to-month, One ye~
## $ Factura_digital      <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Ye~
```

```
## $ Servicio_Internet     <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic, Fi~
## $ Soporte_tecnico       <fct> No, No, No, Yes, No, No, No, No, Yes, No, No, No~
## $ CopiaSeguridad_Online <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, No, ~
## $ Television_carta      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No, N~
## $ Meses_alta            <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, ~
```

Nos hemos asegurado que cada columna tiene el tipo de dato adecuado, factores (fact) para las columnas categorías o lógicas, y numérico (int) para aquellas que son númericas, en este caso "Meses_alta".

A continuación realizamos un breve Análisis Exploratorio (EDA), lo que queremos conseguir es verificar la integridad de los datos, por ejemplo comprobamos si hay algún null.

```
##  Abandono             Contrato    Factura_digital   Servicio_Internet
##  No :5163   Month-to-month:3875   No :2864        DSL        :2416
##  Yes:1869   One year      :1472   Yes:4168        Fiber optic:3096
##             Two year      :1685                   No         :1520
##
##
##
##             Soporte_tecnico         CopiaSeguridad_Online
##  No                 :3472   No                 :3087
##  No internet service:1520   No internet service:1520
##  Yes                :2040   Yes                :2425
##
##
##
##             Television_carta   Meses_alta
##  No                 :2809    Min.   : 1.00
##  No internet service:1520    1st Qu.: 9.00
##  Yes                :2703    Median :29.00
##                              Mean   :32.42
##                              3rd Qu.:55.00
##                              Max.   :72.00
```

```
##              Abandono              Contrato       Factura_digital
##                     0                     0                     0
##     Servicio_Internet       Soporte_tecnico CopiaSeguridad_Online
##                     0                     0                     0
##      Television_carta            Meses_alta
##                     0                     0
```

Verificamos que no existen nulos en ninguna de las columnas, por lo que no es necesario hacer más limpieza.

En la actividad se nos pide que hagamos el árbol de decisión basándonos en estas columnas exclusivamente: Contrato, Factura digital, Servicio Internet, Soporte técnico, Copia de Seguridad Online, Televisión, Meses de alta en el servicio.

Sin embargo, queremos antes llevar a cabo dos o tres hipótesis por si encontráramos algunas correlaciones espurias o anomalías entre la variable de estudio "Abandono" y algunas otras variables.

Hipótesis

Hipótesis 1: La gente con contrato mensual tiende a abandonar más el servicio.

```
##
##                          No       Yes
##   Month-to-month 57.290323 42.709677
```

```
##   One year       88.722826 11.277174
##   Two year       97.151335  2.848665
```

Descartamos la Hipótesis Nula en este caso. Parece haber una alta probabilidad de que en el caso de los contratos pagagos mes a mes la tasa de abandono sea mayor

Hipotesis 2: Los clientes sin soporte técnico son más propensos a irse.

```
##
##                              No       Yes
##   No                  58.352535 41.647465
##   No internet service 92.565789  7.434211
##   Yes                 84.803922 15.196078
```

De nuevo, descartamos la hipotesis nula: hay probabilidad alta de marcharse cuando el cliente no tienen soporte técnico.

conclusiones de las hipotesis:

Parece haber diferencias probabables significativas según las variables independientes y la variable de estudio. Por lo que procedemos a ejecutar el modelo de clasificación

Modelado: Regresión Logística

Establecemos una semilla (123) para asegurar reproducibilidad

Partición de Datos (Train/Test)

Realizamos una partición estratificada 80/20 para asegurar la representatividad de la clase objetivo.

Ajuste del Modelo Logit

Utilizamos un modelo lineal generalizado (glm) con familia binomial.

```
##
## Call:
## glm(formula = Abandono ~ Contrato + Factura_digital + Servicio_Internet +
##     Soporte_tecnico + CopiaSeguridad_Online + Television_carta +
##     Meses_alta, family = "binomial", data = entrenamiento)
##
## Coefficients: (3 not defined because of singularities)
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -0.525708   0.089465  -5.876 4.20e-09
## ContratoOne year                        -0.797248   0.116075  -6.868 6.49e-12
## ContratoTwo year                        -1.583573   0.193008  -8.205 2.31e-16
## Factura_digitalYes                       0.436468   0.081214   5.374 7.69e-08
## Servicio_InternetFiber optic             0.965997   0.084200  11.473  < 2e-16
## Servicio_InternetNo                     -0.994004   0.139396  -7.131 9.98e-13
## Soporte_tecnicoNo internet service             NA         NA      NA       NA
## Soporte_tecnicoYes                      -0.378429   0.093188  -4.061 4.89e-05
## CopiaSeguridad_OnlineNo internet service        NA         NA      NA       NA
## CopiaSeguridad_OnlineYes                -0.105591   0.084457  -1.250    0.211
## Television_cartaNo internet service             NA         NA      NA       NA
## Television_cartaYes                      0.462113   0.082514   5.600 2.14e-08
## Meses_alta                              -0.032580   0.002341 -13.919  < 2e-16
##
```

```
## (Intercept)                              ***
## ContratoOne year                         ***
## ContratoTwo year                         ***
## Factura_digitalYes                       ***
## Servicio_InternetFiber optic             ***
## Servicio_InternetNo                      ***
## Soporte_tecnicoNo internet service
## Soporte_tecnicoYes                       ***
## CopiaSeguridad_OnlineNo internet service
## CopiaSeguridad_OnlineYes
## Television_cartaNo internet service
## Television_cartaYes                      ***
## Meses_alta                               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6517.2  on 5626  degrees of freedom
## Residual deviance: 4776.9  on 5617  degrees of freedom
## AIC: 4796.9
##
## Number of Fisher Scoring iterations: 6
```

Observamos que variables como CopiaSeguridad_Online tienen un p-valor alto (> 0.05), lo que sugiere, aparentemente, que no son estadísticamente significativas. Sin embargo, sospechamos que existe multicolinealidad con 'Servicio_Internet' (técnicamente redundantes porque si no hay servicio de internet no puede haber servicio de Copia de Seguridad). Al eliminarla, simplificamos el modelo reduciendo el ruido, asumiendo que el efecto de la conectividad ya está capturado por la variable de Internet.

Procedemos a refinar el modelo eliminando esa variable. También calcularemos el Pseudo R2 para medir la certidumbre alcanzada con nuestro modelo

```
## fitting null model for pseudo-r2
##  McFadden
## 0.2667851
```

```
##
## Call:
## glm(formula = Abandono ~ Contrato + Factura_digital + Servicio_Internet +
##     Soporte_tecnico + Television_carta + Meses_alta, family = "binomial",
##     data = entrenamiento)
##
## Coefficients: (2 not defined because of singularities)
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.540355   0.088685  -6.093 1.11e-09 ***
## ContratoOne year                 -0.798076   0.116044  -6.877 6.10e-12 ***
## ContratoTwo year                 -1.582358   0.192957  -8.201 2.39e-16 ***
## Factura_digitalYes                0.432385   0.081146   5.329 9.90e-08 ***
## Servicio_InternetFiber optic      0.967790   0.084181  11.497  < 2e-16 ***
## Servicio_InternetNo              -0.967986   0.137897  -7.020 2.22e-12 ***
## Soporte_tecnicoNo internet service       NA         NA      NA       NA
## Soporte_tecnicoYes               -0.382357   0.093119  -4.106 4.02e-05 ***
## Television_cartaNo internet service      NA         NA      NA       NA
## Television_cartaYes               0.461890   0.082496   5.599 2.16e-08 ***
## Meses_alta                       -0.033413   0.002247 -14.870  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6517.2  on 5626  degrees of freedom
## Residual deviance: 4778.5  on 5618  degrees of freedom
## AIC: 4796.5
##
## Number of Fisher Scoring iterations: 6
```

Evaluación del Logit

Calculamos la matriz de confusión y el Accuracy sobre el conjunto de Test. Utilizamos un umbral de decisión del 50%

```
##      prediccion_clase
##        No Yes
##   No  927 105
##   Yes 181 192
```

```
## [1] "Accuracy Logit (Umbral 0.5): 79.64 %"
```

Este modelo vale para predecir bien aquellos que se quedan, pero no predice bien aquellos que se marchan (Falsos negativos)

Asumimos que nuestro objetivo es retener clientes (evitar que se vayan).

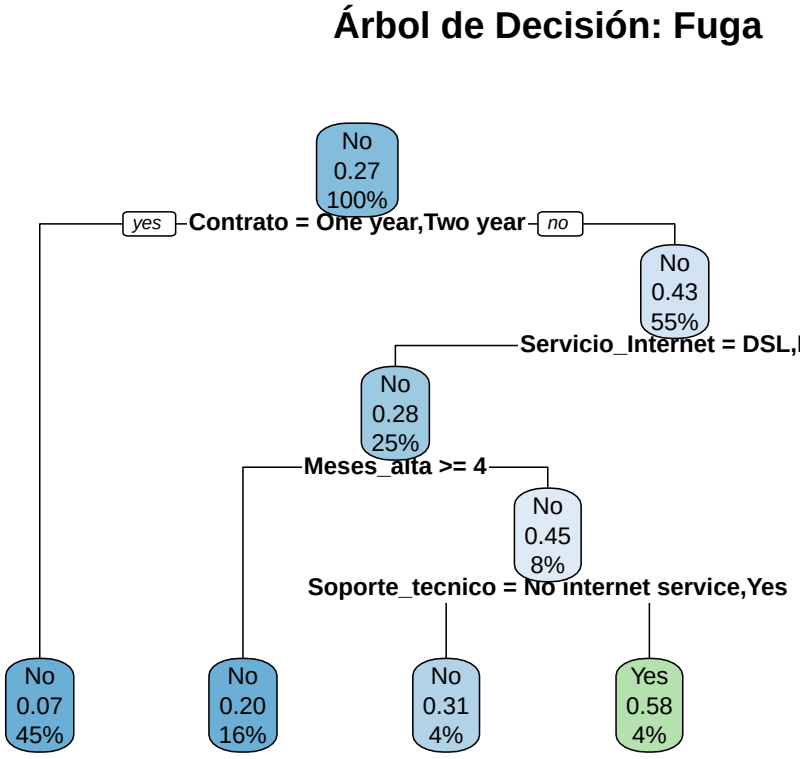
¿Qué error me cuesta más dinero? Al bajar el umbral a 0.30, sacrificamos Accuracy global y Especificidad (aumentan las falsas alarmas), pero aumentamos drásticamente la Sensibilidad (Recall). Esto alinea el modelo con el objetivo de negocio: es más barato incentivar a un cliente que se iba a quedar (Falso Positivo) que perder a un cliente real por no detectarlo (Falso Negativo).

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  775  95
##        Yes 257 278
##
##                Accuracy : 0.7495
##                  95% CI : (0.7259, 0.7719)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 0.1073
##
##                   Kappa : 0.4358
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7453
##             Specificity : 0.7510
##          Pos Pred Value : 0.5196
##          Neg Pred Value : 0.8908
##              Prevalence : 0.2655
##          Detection Rate : 0.1979
##    Detection Prevalence : 0.3808
```

```
##       Balanced Accuracy : 0.7481
##
##        'Positive' Class : Yes
##
```

```
## [1] "Sensibilidad (Recall) con umbral 0.3: 0.7453"
```

5. Modelado: Árbol de Decisión



Hacemos arbol de decisión para comprar modelos

Evaluamos el modelo del arbol

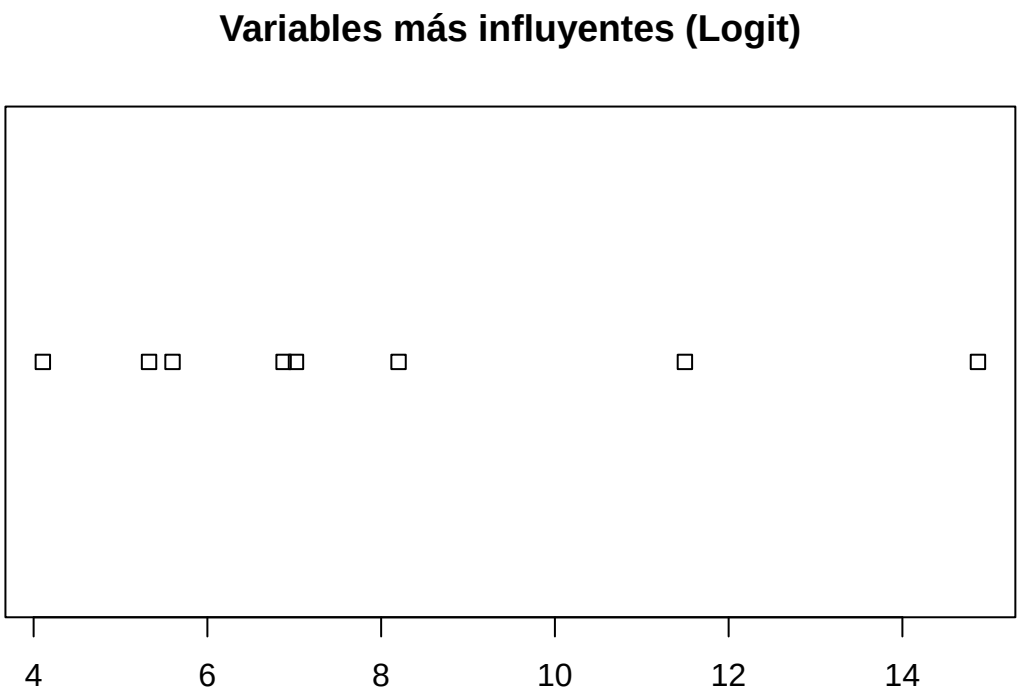
```
##      prediccion_arbol
##        No Yes
##   No  942  90
##   Yes 199 174
```

```
## [1] "Accuracy Árbol: 79.43 %"
```

El arbol muestra un modelo de toma de decisiones peor para nuestro objetivo de negocio de retención del clientes y para el error que cuesta más dinero.

conclusiones y comparativa

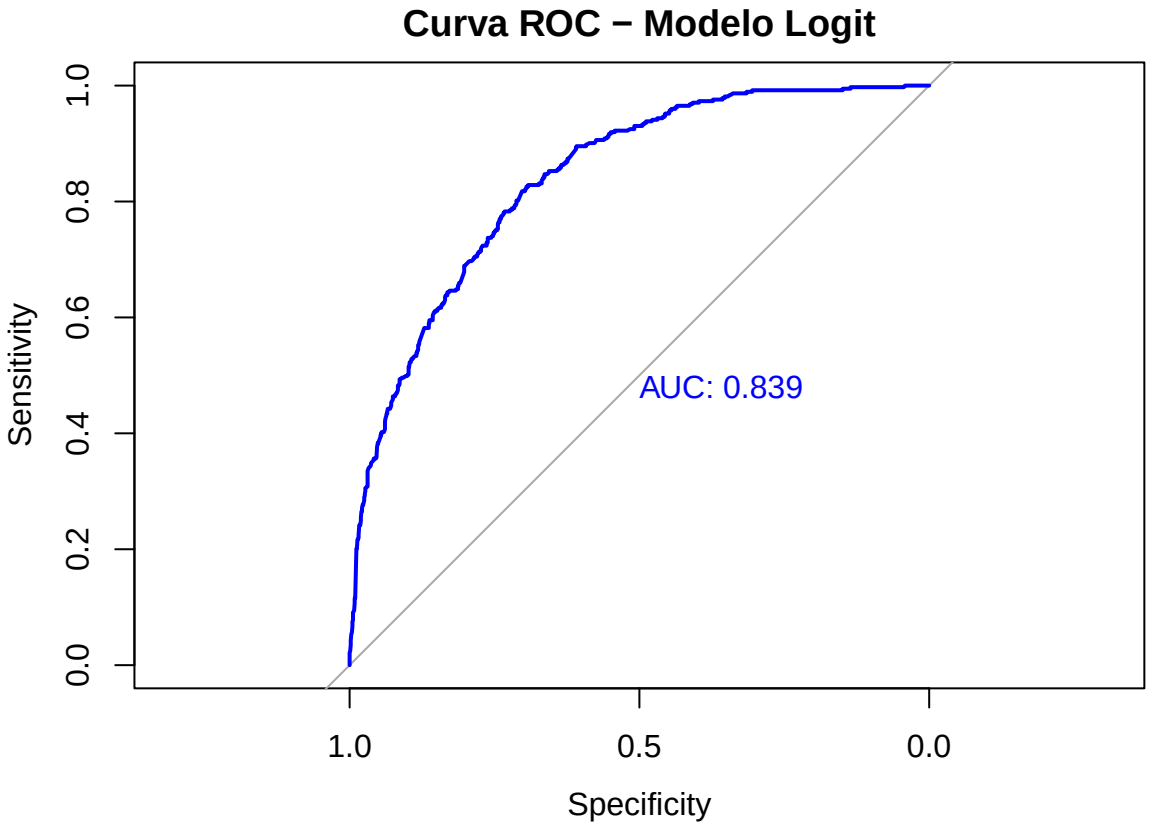
Importancia de las variables




```
## NULL
```

```
##                                Overall
## Meses_alta                   14.869585
## Servicio_InternetFiber optic 11.496550
## ContratoTwo year              8.200584
## Servicio_InternetNo           7.019640
## ContratoOne year              6.877384
## Television_cartaYes           5.598935
## Factura_digitalYes            5.328508
## Soporte_tecnicoYes            4.106108
```

Curva ROC



Conclusión Final

El modelo Logit (ajustado a umbral 0.3) presenta un Accuracy del 74.9%, mientras que el Árbol obtiene un 79.4%.

Aunque el Accuracy del árbol pueda parecer superior en ciertos contextos, basándonos en el AUC y la capacidad de ajustar el umbral para priorizar la Sensibilidad (Recall), recomendamos el modelo Logit. Este nos permite cuantificar mejor cuánto aumenta el riesgo de fuga (Odds Ratio) y adaptar la estrategia de retención para minimizar la pérdida de clientes reales, aunque pensamos que sería demasiado arriesgado implementarlo en un entorno real.