MODELO LINEAL GENERALIZADO:

En primer lugar se realizó un **modelo lineal generalizado** tomando como variable dependiente churn, y las demás como independientes:

```
#GENERALIZED LINEAR MODEL:
GLM.1 <- glm(churn ~ antiguedad_anos + canal_digital +cod_cliente + edad +
otros_ramos + prima_anual_eur +
  siniestro_ultimo_ano, family=binomial(logit), data=Dataset)
summary(GLM.1)
exp(coef(GLM.1))  # Exponentiated coefficients ("odds ratios")
```

```
Call:
glm(formula = churn ~ antiguedad_anos + canal_digital + cod_cliente + edad +
otros_ramos + prima_anual_eur + siniestro_ultimo_ano,
    family = binomial(logit), data = Dataset)

Coefficients:
                         Estimate   Std. Error z value Pr(>|z|)
(Intercept)          -3.065689377  0.132110725 -23.205  < 2e-16 ***
antiguedad_anos      -0.075340949  0.008838240  -8.524  < 2e-16 ***
canal_digital         0.472047594  0.070362313   6.709 1.96e-11 ***
cod_cliente          -0.000002107  0.000001785  -1.181    0.238
edad                 -0.013253541  0.001911528  -6.933 4.11e-12 ***
otros_ramos          -0.624294793  0.069042217  -9.042  < 2e-16 ***
prima_anual_eur       0.000531816  0.000063347   8.395  < 2e-16 ***
siniestro_ultimo_ano  0.968313575  0.073876448  13.107  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En las estimaciones de los coeficientes vemos que todas las variables menos `cod_cliente` son significativas, lo que tiene sentido porque código cliente es simplemente la enumeración de 1 a 50.000 para identificar a los clientes.

En concreto, destacan `canal_digital`, con una relación positiva con la variable dependiente (β=0.47), `otros_ramos` con una relación negativa (β=-0.624), y en especial `siniestro_ultimo_año`, con una relación positiva de β=0.96.

Por general, la probabilidad de baja aumenta con el aumento de `canal_digital`, `prima_anual_eur` y `siniestro_ultimo_año`, y disminuye con aumento en `antiguedad_años`, `edad` y `otros_ramos`.

Respecto al hecho de que `prima_anual_eur` parezca que tenga un efecto tan pequeño comparado con otras variables, hay que recordar que el coeficiente (β) indica el cambio esperado en la variable dependiente cuando la variable explicativa aumenta una unidad, manteniendo las demás constantes. En un modelo logístico, los coeficientes representan **el cambio en el logaritmo de las probabilidades (log-odds)** de que ocurra el evento (`churn` = *1*) **por cada unidad adicional de la variable**. Por tanto, el valor del coeficiente depende de las unidades de medida de esa variable. En este caso, por cada 1 euro adicional en la prima anual, el *logit* (log de las odds de `churn`) aumenta en 0.000531816. Aun así, veremos más claramente los efectos de cada variable explicativa con el odds ratio.

```
> exp(coef(GLM.1))  # Exponentiated coefficients ("odds ratios"
(Intercept)   antiguedad_anos   canal_digital   cod_cliente   edad
0.04662169    0.92742723        1.60327369      0.99999789    0.98683390
otros_ramos      prima_anual_eur      siniestro_ultimo_ano
0.53563903       1.00053196           2.63349951
```

El **odds ratio**, que se calcula como e^{β}, cuantifica la **fuerza y dirección de la asociación** entre una variable explicativa y la probabilidad de que ocurra un determinado evento. Veíamos que si OR = 1 (β=0), la variable no influye en la probabilidad de ocurrencia del evento. Si es menor que 1 (β<0), reduce dicha probabilidad de ocurrencia, y si es mayor que 1 (β>0), la aumenta.

Por tanto, la dirección en que afectan a la variable dependiente se mantiene igual a como vimos con las betas, pero ahora podemos interpretar y cuantificar más claramente el efecto:

• `antiguedad_años`: 0.92742723 → por cada año extra de antigüedad, la probabilidad de baja disminuye un 7.26%.

• `canal_digital`: 1.60327369 → esta variable es binomial, por lo que la interpretación en este caso sería que los clientes que utilizan el canal digital tienen 60% más de probabilidades de darse de baja que los que no.

• `cod_cliente`: 0.99999789 → en este caso, el modelo ha visto que a mayor posición que tiene el cliente en esta lista, menos probabilidad de churn, pero recordemos que esa beta no era significativa (se trata únicamente de un identificador del cliente por lo que se debería eliminar esta variable del modelo).

• `edad`: 0.98683390 → por cada año adicional de edad, la probabilidad de churn disminuye un 1.32%.

• `otros_ramos`: 0.53563903 → de nuevo es variable binomial. Indica que si el cliente tiene otros ramos (otros productos o servicios contratados), la probabilidad de churn disminuye 46.43%.

• `prima_anual_eur`: 1.00053196 → aquí como vimos antes debemos tener cuidado al interpretar ya que parecería a simple vista que el efecto de esta variable no tiene especial importancia, pero es debido a las unidades en que está expresada. Vemos que significa que por cada euro adicional que el cliente paga de prima anual, la probabilidad de churn aumenta un 0.053%. Pero los cambios de 1 euro no son significativos, por lo que lo mejor es interpretarlo en intervalos más realistas, por ejemplo, 100 €. De manera que, por cada incremento en la prima de 100€, la probabilidad de churn se incrementa en un 5.5%, lo que ya parece un efecto más relevante.

• `siniestro_ultimo_año`: 2.63349951 → es variable binaria. Indica que si el cliente ha tenido un siniestro en el último año, la probabilidad de churn son 2.63 veces mayores que los que no lo tuvieron, lo que equivale a un incremento del 163% en el riesgo de baja. Es uno de los efectos más fuertes del modelo.

(En todos los casos es *Ceteris Paribus*)

Los resultados del modelo logístico muestran un **patrón coherente con el comportamiento esperado de los clientes**:

Los clientes con mayor antigüedad y de mayor edad presentan una menor probabilidad de causar baja, lo que sugiere una mayor estabilidad y fidelización. Además, aquellos que tienen otros ramos contratados (es decir, más de un producto o servicio) también muestran una menor propensión al churn, reforzando la idea de que la diversificación de productos favorece la retención.

Por el contrario, el uso del canal digital se asocia con una mayor probabilidad de abandono, lo que podría reflejar un perfil de cliente más independiente o sensible a las comparativas de precios. Un incremento en la prima anual aumenta ligeramente la probabilidad de baja, posiblemente por percepción de sobrecoste o falta de valor percibido. Y haber tenido un siniestro durante el último año se relaciona con un riesgo de baja significativamente superior, lo que sugiere que las experiencias recientes con la compañía influyen fuertemente en la decisión de continuar o no.

Por tanto, el modelo proporciona una **base sólida para poder diseñar estrategias de retención focalizadas en los segmentos de mayor riesgo**. Así, se proponen **líneas de actuación** para reducir el *churn* y aumentar la retención:

- En primer lugar, se recomienda **reforzar la fidelización** mediante la vinculación de productos o servicios adicionales. Ofrecer beneficios por varias contrataciones, programas de puntos o paquetes personalizados podría consolidar este efecto de que los clientes con mayor número de contratos tienden a mantener una relación más estable.
- En cuanto al canal digital, convendría **mejorar la propuesta de valor y la diferenciación** frente a la competencia. Dado que estos clientes pueden cambiar fácilmente de compañía, se sugiere potenciar la experiencia digital con servicios exclusivos, atención más cercana o ventajas personalizadas que fomenten la permanencia, incluso en un entorno de alta comparabilidad.
- Respecto a la prima anual, se recomienda **analizar las revisiones de precio**, valorando de forma **individualizada** si los incrementos de prima compensan el posible abandono, considerando el perfil del cliente, su antigüedad, el nivel de siniestralidad y su valor a largo plazo. Una política de precios segmentada y basada en valor, junto con incentivos o bonificaciones por permanencia, podría reducir el impacto negativo de los incrementos de prima.
- Finalmente, el hallazgo de que los clientes que han tenido siniestros recientes presentan un riesgo significativamente mayor de baja resalta la necesidad de **gestionar adecuadamente la experiencia post-siniestro**. Una comunicación empática, procesos ágiles de resolución y una atención preferente podrían reducir la insatisfacción y transformar una experiencia potencialmente negativa en una oportunidad de retención. Pero, de nuevo, habría que estudiar si esto resulta más rentable que reemplazar a los clientes que abandonan (análisis coste-beneficio).

En conjunto, las medidas deben orientarse hacia una **gestión preventiva** del abandono, basada en la **personalización**, la **segmentación** y la **mejora del valor percibido**, con el objetivo de fortalecer la relación a largo plazo con los clientes más valiosos.

MATRIZ DE CONFUSIÓN:

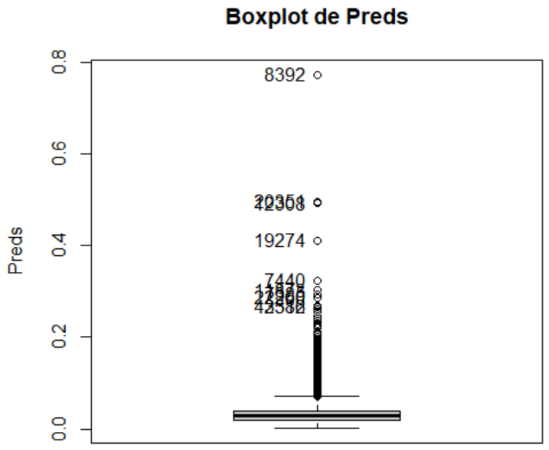
```
#CONFUSSION MATRIX:
Predicciones <- predict(GLM.1, type = "response")
Dataset$Preds <- Predicciones
head(Dataset)
Dataset$Clase_preds <- ifelse(Dataset$Preds >=0.5,1,0)
CM <- table(Dataset$churn, Dataset$Clase_preds)
CM
```

Se ha creado una columna en el Dataset con las predicciones para cada fila (`Preds`), y luego, usando el umbral de 0.5, se generó otra columna en base a la anterior que clasificara en 1 o 0 según fuese >= 0.5 o <0.5 respectivamente (`Clase_preds`).

Con esta columna, `Clase_preds`, y la real, `churn`, se generó la matriz de confusión (MC):

```
          0      1
0    48418      0
1     1581      1
```

Vemos que la matriz está muy desequilibrada, teniendo muchos casos en que se predecía que había baja y efectivamente la había, y muchos en que se decía que habría baja pero no la había (falsos positivos). No hay ningún caso en que se diga que no hay pero la haya, y 1 en que se dice que la hay y la hay. Esto se debe a que el **umbral del 50 %** (0.5) es una convención que solo tiene sentido cuando las **clases están balanceadas**. En nuestro caso las predicciones están claramente desbalanceadas como se ve en el siguiente Boxplot de la variable `Preds`:



Efectivamente, vemos que las probabilidades se concentran en valores bajos, con una media de 0.03164 y std de 0.02087. Por tanto, al fijar el umbral de 0.5 estamos siendo excesivamente conservadores, al estar calificando muchas bajas reales como 'no bajas'.

Con el código `Dataset[Dataset$Preds >= 0.5, ]`, se observa que solo una observación (la fila 8392) supera ese umbral, correspondiendo precisamente al único caso correctamente clasificado como churn en la matriz de confusión inicial (1/1).

Por otro lado, el uso del umbral 0.5 es apropiado únicamente cuando los **costes de error** (falsos positivos y falsos negativos) son similares. En este caso, sin embargo, se podría considerar **más costoso un falso negativo**, ya que aunque las acciones de retención suponen un coste para la empresa (asociado a los falsos positivos), en la práctica **el coste de los falsos negativos suele ser mayor**, ya que implican la pérdida efectiva de clientes. Por tanto, resulta preferible un modelo más sensible, que minimice las fugas reales aunque incremente el número de falsas alarmas.

Por ello, resulta necesario **ajustar el umbral** para encontrar un equilibrio adecuado entre **sensibilidad** (capacidad para detectar churns reales) y **precisión** (evitar falsas alarmas).

Así, se pensó en fijar el umbral en torno a la **probabilidad base observada de churn:**

`sum(Dataset$churn))/length(Dataset$churn) #= 0.03164 = (1.582/50.000)`

Este 3.164%, que corresponde al porcentaje real de clientes que se dan de baja (1.582 de 50.000) observado en la variable `churn`, permite calibrar el modelo con la frecuencia real del

evento, mejorando el balance entre verdaderos positivos y negativos. Con este umbral, la matriz de confusión queda como sigue:

```
          0     1
0     28632 19786
1       570  1012
```

Efectivamente, en esta matriz vemos que hay un **reparto más equilibrado** entre las secciones. Ahora vemos más casos en que se dice que se va cuando no se va (falsos positivos), y casos en que se dice que no se va y se va (falsos negativos). Por tanto, el modelo se ha vuelto **más sensible** que antes, logrando captar mejor a los clientes potencialmente en riesgo, lo que permite enfocar con mayor precisión nuestros recursos en la retención de esos casos con mayor probabilidad de abandono. A su vez el aumento en falsos positivos implican gasto adicional en acciones de retención, pero este intercambio entre sensibilidad y precisión es esperable y aceptable en una primera calibración del modelo.

En consecuencia, aunque esta no sea todavía la matriz de confusión óptima, **representa una buena aproximación inicial**, que equilibra razonablemente la detección de bajas reales con un nivel de error manejable.

Aun así, como se ha mencionado, este umbral se deberá ajustar. En este caso, dado que las clases están muy desbalanceadas y la mayoría de probabilidades son bajas, es lógico optar por un umbral inferior al 0.5. Sin embargo, la **elección definitiva** del umbral deberá basarse en un **análisis coste-beneficio** en que se compare el coste de falsos positivos VS falsos negativos, evaluando qué tipo de error resulta más caro para la empresa: no detectar a un cliente que se va, o invertir recursos innecesariamente en retener a uno que no se iba a marchar. Según ese análisis, se fijará el umbral:

- Más bajo → menos falsos negativos.
- Más altos → menos falsos positivos.

CONCLUSIÓN:

En síntesis, el modelo logístico desarrollado permite identificar los principales factores asociados al churn y estimar con razonable precisión la probabilidad de baja de cada cliente.

La calibración del umbral en torno a la tasa real de abandono ha mejorado la sensibilidad del modelo, reduciendo las fugas no detectadas a costa de un incremento asumible de falsas alarmas, lo que es adecuado para problemas donde el coste de perder un cliente es superior al de aplicar acciones preventivas sobre clientes que finalmente permanecen.

En conjunto, el análisis proporciona una herramienta útil para orientar **estrategias de retención más efectivas y basadas en datos**, optimizando la gestión del riesgo de abandono y la rentabilidad a largo plazo.