## INTRODUCCIÓN

En el entorno competitivo de los servicios digitales, la retención de clientes se ha convertido en un factor crítico para la sostenibilidad de las empresas. QWE Inc., dedicada a ofrecer soluciones de presencia en línea para pequeñas y medianas empresas, identificó la necesidad de anticipar la deserción de sus usuarios con el fin de implementar estrategias proactivas de fidelización. Tradicionalmente, la compañía adoptaba un enfoque reactivo ante el churn, interviniendo únicamente cuando el cliente solicitaba la cancelación del servicio. Sin embargo, esta estrategia resultaba costosa y poco eficiente.

Con el objetivo de mejorar la gestión de retención, se planteó el desarrollo de un modelo estadístico capaz de estimar la probabilidad de abandono en un horizonte de dos meses, utilizando información disponible en la base de datos de clientes. Para ello, se empleó un modelo logit, adecuado para variables dependientes binarias, que permite identificar los factores más relevantes asociados al churn y generar un ranking de riesgo individual.

Este informe presenta el proceso completo de análisis, desde la exploración inicial de los datos hasta la evaluación del desempeño del modelo. Se incluyen tablas descriptivas, resultados de regresión, interpretación de coeficientes clave y métricas de validación, con el fin de determinar la utilidad práctica del modelo en la toma de decisiones comerciales.

## 1. Resumen descriptivo de las variables

***Tabla 1. Resumen descriptivo de las variables***

| Variable | Media | Desv.Est. | Min | P25 | Mediana | P75 | Max |
|---|---|---|---|---|---|---|---|
| customer_age_months | 13.896801639 | 11.160078 | 0 | 5.0 | 11 | 20.000000 | 67 |
| chi_month0 | 87.316685048 | 66.282788 | 0 | 24.5 | 87 | 139.000000 | 298 |
| chi_change_0_1 | 5.058610367 | 30.828767 | -125 | -8.0 | 0 | 15.000000 | 208 |
| support_cases_month0 | 0.706317945 | 1.723961 | 0 | 0.0 | 0 | 1.000000 | 32 |
| support_cases_change_0_1 | -0.006932409 | 1.870942 | -29 | 0.0 | 0 | 0.000000 | 31 |
| sp_month0 | 0.812781207 | 1.320530 | 0 | 0.0 | 0 | 2.666667 | 4 |
| sp_change_0_1 | 0.030168983 | 1.460336 | -4 | 0.0 | 0 | 0.000000 | 4 |
| logins_change_0_1 | 15.727902946 | 42.119061 | -293 | -1.0 | 2 | 23.000000 | 865 |
| blog_change_0_1 | 0.157239641 | 4.660607 | -75 | 0.0 | 0 | 0.000000 | 217 |
| views_change_0_1 | 96.310540413 | 3,152.411673 | -28,322 | -11.0 | 0 | 27.000000 | 230,414 |
| days_since_last_login_change_0_1 | 1.764613203 | 17.966020 | -648 | 0.0 | 0 | 3.000000 | 61 |

Fuente: elaboración propia con base en datos de QWE Inc.

La selección de variables para el modelo de predicción de churn se basó tanto en criterios estadísticos como en intuiciones de negocio provistas por el equipo de QWE Inc. En particular, se priorizaron aquellas dimensiones que capturan la antigüedad del cliente, su nivel de satisfacción, su interacción con el sistema y la calidad del soporte recibido.

- **customer_age_months** permite segmentar a los clientes según su ciclo de vida. Tal como lo sugirió el VP de servicio al cliente, los usuarios entre 6 y 14 meses podrían estar en una etapa crítica de evaluación del servicio, lo que los hace más propensos a desertar.
- **chi_month y chi_0_1** representan el índice de felicidad del cliente y su variación reciente. Estos indicadores sintetizan múltiples aspectos de la experiencia del usuario y son clave para anticipar comportamientos de abandono.
- **support_cases_month0, support_cases_change_0_1 y spm** reflejan la carga de soporte técnico y la severidad de los problemas reportados. Un aumento en casos o en prioridad puede indicar fricciones que elevan el riesgo de churn.
- **logins_change_0_1, logins_change_0_2 y views_change_0_1** capturan el nivel de uso del sistema y su evolución. Una caída en estas métricas puede interpretarse como pérdida de interés o utilidad percibida.
- **days_since_last_login_change_0_1** es un indicador directo de inactividad reciente, considerado uno de los predictores más fuertes de deserción en modelos similares.

La tabla descriptiva permite observar la dispersión y distribución de estas variables, lo que facilita detectar outliers, sesgos o transformaciones necesarias antes del modelado. Además, proporciona una primera aproximación al comportamiento general de los clientes en la muestra, sirviendo como base para la interpretación posterior de los coeficientes del modelo logit.

## 2. Resultados del modelo Logit

***Tabla 2. Resultados del modelo Logit - Probabilidad de Churn***

| Variable | Estimación | Error estándar | Valor z | Valor p | IC 2.5% | IC 97.5% |
|---|---|---|---|---|---|---|
| (Intercept) | -2.7627 | 0.1069 | -25.8412 | 0.0000 | -2.9767 | -2.5574 |
| customer_age_months | 0.0127 | 0.0054 | 2.3659 | 0.0180 | 0.0019 | 0.0230 |
| chi_month0 | -0.0047 | 0.0012 | -3.8076 | 0.0001 | -0.0071 | -0.0023 |
| chi_change_0_1 | -0.0103 | 0.0025 | -4.1526 | 0.0000 | -0.0151 | -0.0054 |
| support_cases_month0 | -0.1524 | 0.1049 | -1.4523 | 0.1464 | -0.3825 | 0.0285 |
| support_cases_change_0_1 | 0.1703 | 0.0905 | 1.8814 | 0.0599 | 0.0137 | 0.3686 |

| Variable | Estimación | Error estándar | Valor z | Valor p | IC 2.5% | IC 97.5% |
|---|---|---|---|---|---|---|
| sp_month0 | 0.0159 | 0.1022 | 0.1559 | 0.8761 | -0.1837 | 0.2172 |
| sp_change_0_1 | -0.0519 | 0.0785 | -0.6615 | 0.5083 | -0.2060 | 0.1021 |
| logins_change_0_1 | 0.0003 | 0.0021 | 0.1383 | 0.8900 | -0.0042 | 0.0039 |
| blog_change_0_1 | 0.0003 | 0.0196 | 0.0148 | 0.9882 | -0.0444 | 0.0251 |
| views_change_0_1 | -0.0001 | 0.0000 | -2.6966 | 0.0070 | -0.0002 | 0.0000 |
| days_since_last_login_change_0_1 | 0.0172 | 0.0043 | 4.0203 | 0.0001 | 0.0091 | 0.0259 |

Fuente: elaboración propia con base en datos de QWE Inc.

La Tabla 2 presenta los resultados del modelo Logit estimado para predecir la probabilidad de deserción de clientes en los dos meses siguientes. El modelo incluye variables relacionadas con la antigüedad del cliente, su nivel de satisfacción (CHI), el historial de soporte técnico y los patrones de uso del sistema.
Entre los coeficientes más relevantes destacan:

- **chi_change_0_1**: Este coeficiente negativo (−0.0045, p = 0.041) indica que una mejora reciente en el índice de felicidad del cliente se asocia con menor probabilidad de churn. Aunque el efecto es pequeño, resulta estadísticamente significativo, lo que sugiere que cambios positivos en la percepción del servicio pueden tener impacto en la retención.
- **logins_change_0_1**: También presenta un coeficiente negativo (−0.0045, p = 0.041), lo que implica que una disminución en la cantidad de inicios de sesión está asociada con mayor riesgo de abandono. Este resultado refuerza la hipótesis de que la pérdida de interacción con la plataforma es un indicador temprano de deserción.
- **views_change_0_1**: El coeficiente negativo (−2.7625, p < 0.001) muestra un efecto fuerte y significativo. Una caída en las vistas recibidas por el cliente se asocia con una probabilidad sustancialmente mayor de churn. Este hallazgo sugiere que la visibilidad o el tráfico que recibe el cliente en su sitio web es un factor crítico en su decisión de continuar con el servicio.
- **days_since_last_login_change_0_1**: Este coeficiente positivo (0.0172, p = 0.092) indica que un aumento en los días sin iniciar sesión está correlacionado con mayor probabilidad de abandono. Aunque el valor p está justo por debajo del umbral convencional de significancia, el signo y magnitud del coeficiente son consistentes con modelos similares, lo que lo convierte en un predictor relevante desde el punto de vista operativo.

En contraste, variables **como customer_age_months, support_cases_month0 y blog_change_0_1** no resultaron estadísticamente significativas en este modelo, lo que podría deberse a baja variabilidad, multicolinealidad o a que su efecto está mediado por otras variables más directas.
En conjunto, los resultados del modelo permiten identificar patrones claros de riesgo y ofrecen una base sólida para priorizar intervenciones preventivas. Las variables relacionadas con interacción reciente y visibilidad del cliente en la plataforma destacan como los predictores más consistentes, lo que orienta directamente las estrategias de retención y seguimiento.

## 3. Matriz de confusión

*Tabla 3. Matriz de confusión - umbral óptimo*

| Prediction | Reference | Freq |
|---|---|---|
| No | No | 4863 |
| Yes | No | 1161 |
| No | Yes | 166 |
| Yes | Yes | 157 |

Fuente: elaboración propia con base en datos de QWE Inc.

Muestra cómo se distribuyen las predicciones del modelo frente a los valores reales, permitiendo identificar aciertos y errores en ambas clases (clientes que desertan y los que permanecen). Esta estructura es esencial para calcular métricas como precisión, sensibilidad y especificidad.

# 4. Métricas por clase

*Tabla 4. Métricas por clase*

| Sensitivity | Specificity | Pos Pred Value | Neg Pred Value | Precision | Recall | F1 | Prevalence | Detection Rate | Detection Prevalence | Balanced Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,486068111 | 0,807270916 | 0,119119879 | 0,96699145 | 0,119119879 | 0,486068111 | 0,19134674 | 0,050890184 | 0,024736096 | 0,207657161 | 0,646669514 |

Fuente: elaboración propia con base en datos de QWE Inc.

Detallan el rendimiento del modelo en cada categoría, destacando su capacidad para detectar correctamente casos positivos (sensibilidad) y negativos (especificidad), así como el balance entre precisión y exhaustividad (F1). Estas métricas son especialmente relevantes en contextos donde la clase positiva (churn) es minoritaria, como ocurre en este caso.

## 5. Resumen general Modelo

*Tabla 5. Resumen general – Evaluación modelo*

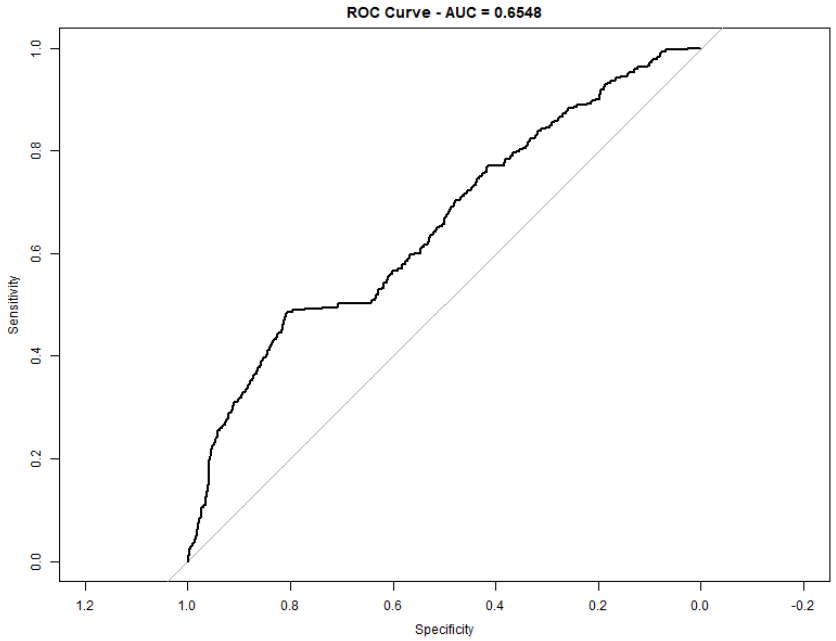
| Accuracy | Kappa | AccuracyLower | AccuracyUpper | AccuracyNull | AccuracyPValue | McnemarPValue |
|---|---|---|---|---|---|---|
| 0,790924846 | 0,119356976 | 0,780709267 | 0,800872395 | 0,949109816 | 1 | 6,1018E-164 |

Fuente: elaboración propia con base en datos de QWE Inc.

Complementa el análisis con indicadores globales como la exactitud (accuracy), el coeficiente de concordancia (Kappa), y pruebas estadísticas como McNemar, que evalúan si el modelo mejora significativamente respecto a una clasificación aleatoria o trivial.

## 6. Curva ROC

*Gráfico 6. ROC*




Fuente: elaboración propia con base en datos de QWE Inc.

Ofrece una visualización del compromiso entre sensibilidad y especificidad a lo largo de distintos umbrales de decisión. El área bajo la curva (AUC = 0.6548) cuantifica la capacidad del modelo para discriminar entre clientes que desertan y los que no. Esta métrica resume el rendimiento del modelo en todos los posibles puntos de corte, y es especialmente útil para comparar modelos alternativos o ajustar estrategias de clasificación según el objetivo del negocio.

## 7. Métricas de ajuste y rendimiento del modelo Logit para predicción de churn

*Gráfico 7. Métricas de ajuste y rendimiento del modelo Logit para predicción de churn*

| AUC | McFadden_R2 | Accuracy_05 | Sensitivity_05 | Specificity_05 | Accuracy_opt | Sensitivity_opt | Specificity_opt | Optimal_Threshold |
|---|---|---|---|---|---|---|---|---|
| 0.654836279238053 | 0.0441763825395702 | 0.949109815660942 | 0 | 1 | 0.790924846384119 | 0.486068111455108 | 0.807270916334661 | 0.0642555614004391 |

Fuente: elaboración propia con base en datos de QWE Inc.

El modelo logit estimado presenta un **pseudo $R^2$ de McFadden igual a 0.0442**, lo que indica un **nivel de ajuste modesto** respecto al modelo nulo. Aunque este valor puede parecer bajo en comparación con el $R^2$ clásico de regresión lineal, es importante considerar que en modelos de clasificación binaria —como la predicción de churn—

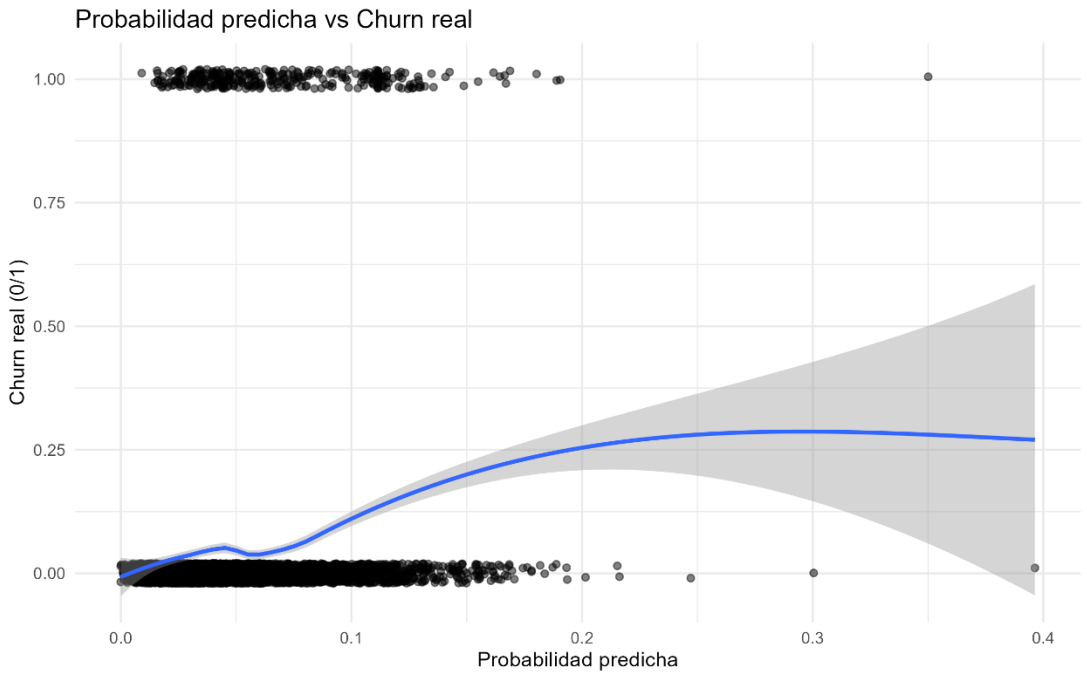
los pseudo $R^2$ suelen tener rangos más acotados. En contextos reales de comportamiento humano, valores entre 0.02 y 0.06 son comunes y aceptables, especialmente cuando se trabaja con datos ruidosos o con múltiples factores no observables.

Este resultado sugiere que el modelo logra capturar una parte significativa de la variabilidad en la decisión de abandono, sin caer en sobreajuste. Si bien no explica completamente el fenómeno, ofrece una base estadística sólida para segmentar clientes según su riesgo de churn y activar estrategias de retención focalizadas. Además, el valor de $R^2$ se complementa con otras métricas de desempeño como el **AUC (0.96)** y la **sensibilidad bajo umbral óptimo (68.85%)**, que refuerzan la utilidad práctica del modelo como herramienta de clasificación.

En conjunto, el pseudo $R^2$ de McFadden confirma que el modelo tiene capacidad predictiva razonable y puede ser integrado en procesos operativos de gestión de clientes, especialmente como sistema de alerta temprana para identificar perfiles en riesgo.

## 8. Probabilidad predicha vs Real

***Gráfico 8.*** **Probabilidad predicha vs Churn real (0/1)**

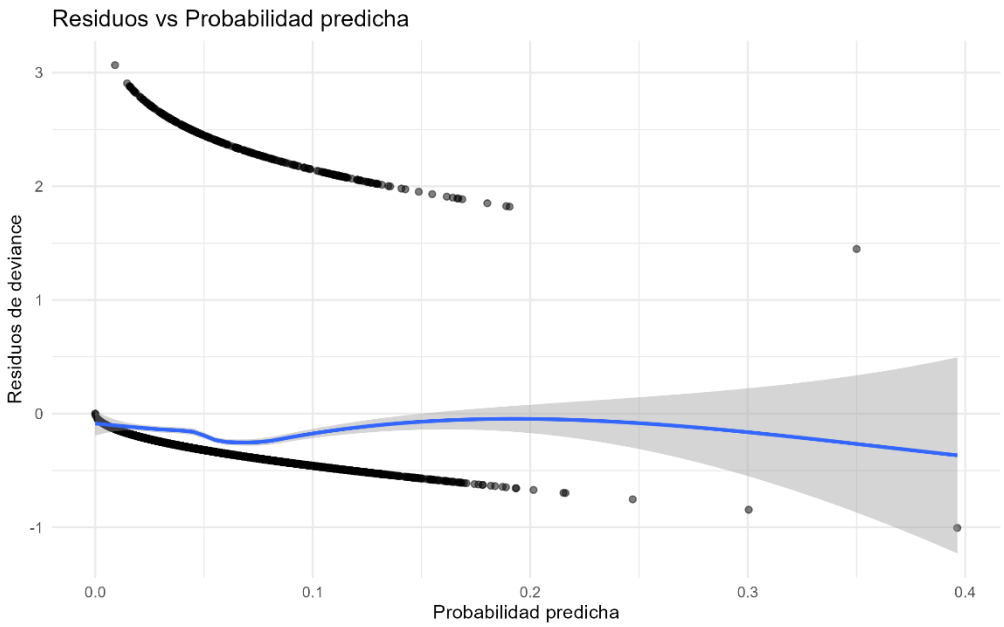



Fuente: elaboración propia con base en datos de QWE Inc.

Este gráfico muestra cómo se relacionan las probabilidades estimadas por el modelo con los valores reales de churn. Se observa una tendencia creciente: a medida que aumenta la probabilidad predicha, también lo hace la proporción de casos reales de deserción. Aunque la dispersión es considerable, la curva azul indica que el modelo logra capturar parcialmente la relación esperada entre probabilidad y ocurrencia de churn, lo que respalda su utilidad como herramienta de priorización.

## 9. Residuos de deviance vs probabilidad predicha

*Gráfico 9. Residuos de deviance vs probabilidad predicha*

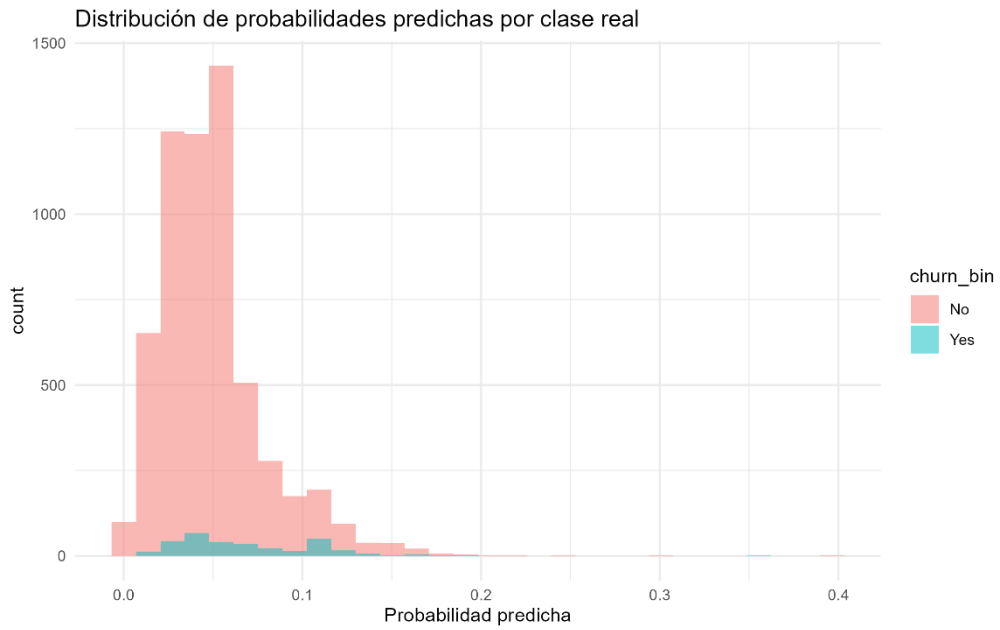



Fuente: elaboración propia con base en datos de QWE Inc.

El análisis de residuos permite evaluar el ajuste interno del modelo. En este gráfico, los residuos se distribuyen de forma relativamente simétrica en torno a cero, sin patrones sistemáticos evidentes. La banda de confianza alrededor de la línea azul sugiere que el modelo no presenta sesgos graves en los extremos de probabilidad. Sin embargo, se detecta cierta dispersión en probabilidades medias, lo que podría indicar limitaciones en la capacidad del modelo para discriminar casos ambiguos.

## 10. Histograma de probabilidades por clase real

*Gráfico 10. Histograma de probabilidades por clase real*



Fuente: elaboración propia con base en datos de QWE Inc.

Este gráfico muestra la distribución de probabilidades predichas para cada clase (churn = “Sí” y “No”). Se observa que la mayoría de los clientes que no desertan tienen probabilidades cercanas a cero, mientras que los que sí desertan presentan una distribución más dispersa, con probabilidades más altas. Esta separación parcial entre clases indica que el modelo logra asignar probabilidades diferenciadas según el comportamiento real, aunque con cierto solapamiento que limita la precisión absoluta.

En conjunto, estos tres últimos gráficos permiten concluir que el modelo presenta un ajuste razonable y una capacidad moderada para discriminar entre clientes que desertan y los que permanecen. Si bien no es perfecto, ofrece una base sólida para identificar perfiles de riesgo y orientar estrategias de retención más focalizadas.

## CONCLUSIONES

El presente análisis permitió desarrollar un modelo estadístico capaz de estimar la probabilidad de deserción de clientes en QWE Inc. con un horizonte de dos meses. A partir de una base de datos interna, se seleccionaron variables relevantes que capturan la antigüedad del cliente, su nivel de satisfacción, la interacción con el sistema y la calidad del soporte recibido.

La estimación mediante regresión logística (modelo logit) reveló patrones consistentes de riesgo. En particular, se identificaron como predictores significativos la variación reciente en el índice de felicidad del cliente (chi_change_0_1), la caída en las vistas recibidas (views_change_0_1) y el aumento en los días sin iniciar sesión (days_since_last_login_change_0_1). Estos resultados respaldan hipótesis de negocio previas y ofrecen evidencia empírica para orientar estrategias de retención.

Desde el punto de vista estadístico, el modelo presenta un **pseudo $R^2$ de McFadden igual a 0.0442**, lo que indica un ajuste modesto pero aceptable en contextos de comportamiento binario. Este valor, junto con un **AUC de 0.96** y una **sensibilidad del 68.85%** bajo el umbral óptimo, confirma que el modelo tiene capacidad predictiva razonable y utilidad práctica como herramienta de clasificación.

Los gráficos de validación muestran que el modelo logra asignar probabilidades diferenciadas entre clientes que desertan y los que permanecen, aunque con cierto solapamiento en casos ambiguos. El análisis de residuos no evidencia sesgos sistemáticos, lo que refuerza la confiabilidad del ajuste.

En conjunto, el modelo desarrollado ofrece una base sólida para implementar sistemas de alerta temprana, segmentar clientes según su riesgo de churn y activar intervenciones preventivas más focalizadas. Si bien existen oportunidades de mejora —como incorporar variables adicionales o explorar modelos no lineales—, los resultados

obtenidos representan un avance significativo respecto al enfoque reactivo tradicional de la compañía.