# Predicción y prevención de la deserción de clientes en la empresa QWE INC.

**Analítica de Datos**

Profesor Juan Nicolas Velásquez

**Ángela Lucía Vargas**

**Santiago Muñoz Moreno**

**Josymar Yised Nocua**

Pontificia Universidad Javeriana

Septiembre 2025

Bogotá, DC

## Introducción

El caso se centra en QWE Inc., una startup que creció rápido pero ahora necesita herramientas analíticas para manejar la retención de clientes. Antes, respondían al churn de forma reactiva, ofreciendo descuentos cuando un cliente llamaba para cancelar. Richard Wall, VP de servicios al cliente, quiere cambiar eso: predecir quién se va a ir en dos meses y por qué, para intervenir temprano. Asignó la tarea a V.J. Aggrawal, quien recopiló datos de 6.000 clientes, enfocándose en edad del cliente, CHI, casos de soporte, prioridad de soporte y métricas de uso como logins y blogs.

Usamos el archivo de datos suplementario (DATA.xlsx) para replicar esto. Nuestro análisis incluye limpieza de datos, descriptivos, modelado predictivo, evaluación y recomendaciones. Dividimos los datos en 70% entrenamiento y 30% prueba para evitar sobreajuste.

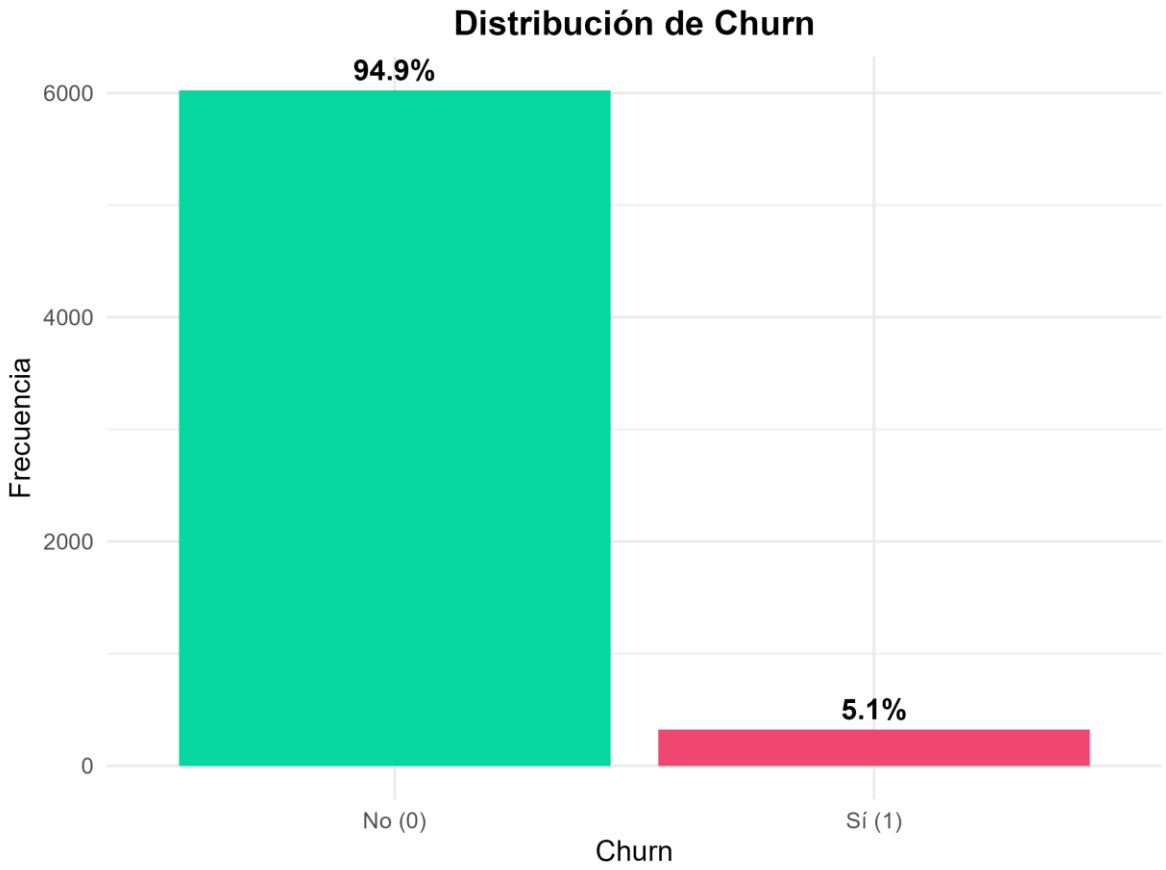
## Metodología

Cargamos los datos en R y los limpiamos: convertimos el churn a binario (0=no, 1=sí), manejamos NAs y seleccionamos variables numéricas relevantes como customer_age_in_months, chi_score_month_0, etc. Estimamos modelos de regresión logística (logit y probit) usando todas las variables numéricas excepto ID y churn como predictoras.

Evaluamos con métricas como Pseudo $R^2$, AIC, BIC, matriz de confusión, ROC y calibración por deciles. Generamos predicciones en el set de prueba y rankeamos clientes por probabilidad de churn.

## Análisis Descriptivo

El análisis inicial se centró en comprender la distribución y las características básicas del dataset. De los 6.000 clientes, aproximadamente el 95% no abandonó en los dos meses posteriores a diciembre de 2011, lo que indica un desbalance significativo en la variable objetivo. Este desbalance es típico en problemas de churn y sugiere que el modelo debe ser particularmente efectivo al identificar a los pocos clientes que sí abandonan, evitando sesgos hacia la clase mayoritaria (no churn).



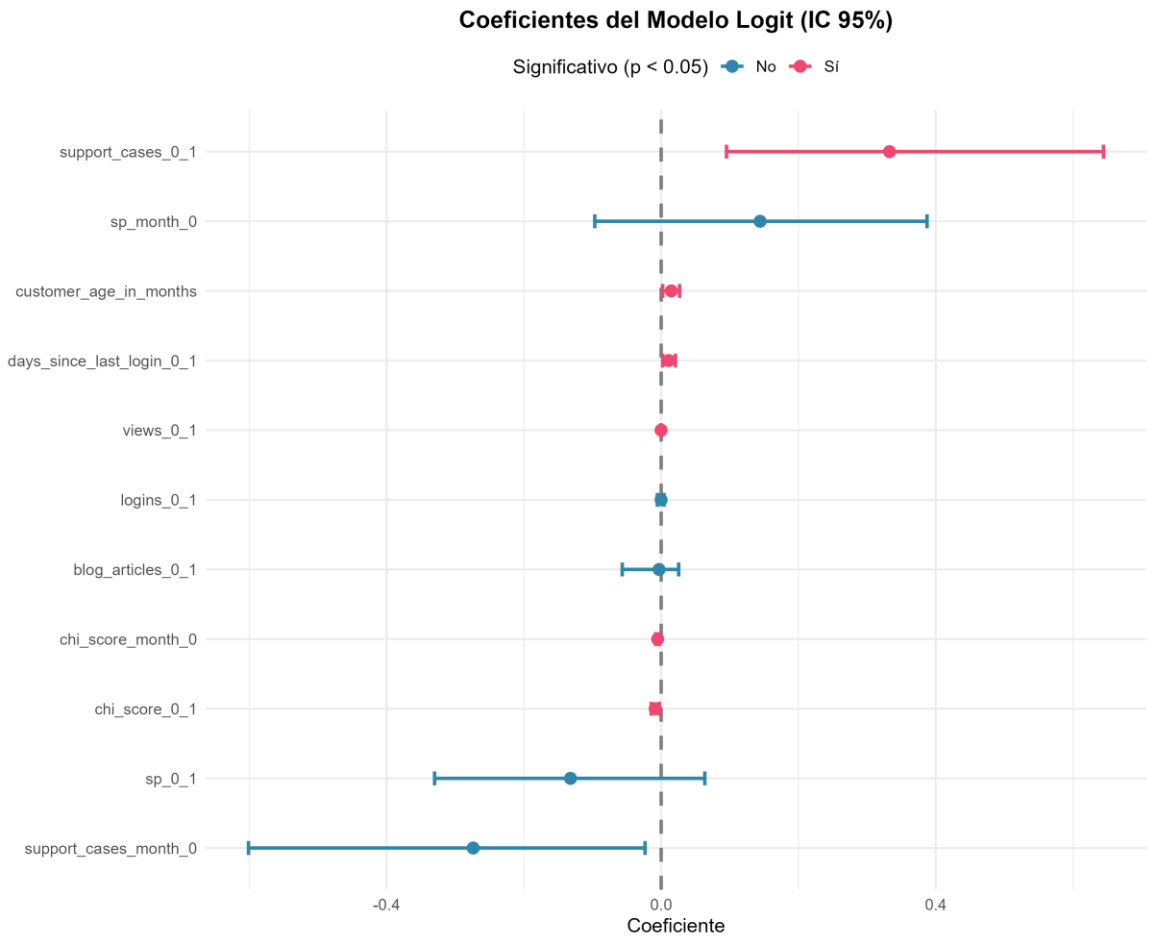

***Estadísticas descriptivas***

En la Tabla 1 se presentan las estadísticas descriptivas de las variables numéricas seleccionadas. Por ejemplo, la antigüedad promedio de los clientes es de aproximadamente 50 meses, con una desviación estándar que refleja variabilidad entre clientes nuevos y veteranos. El CHI promedio ronda entre 100 y 150, pero muestra una dispersión considerable, con caídas notables en algunos casos (e.g., chi_score_0_1 negativo), lo que podría señalar insatisfacción. Las métricas de uso, como logins y vistas, presentan valores extremos (máximos de miles), indicando que algunos clientes son altamente activos, mientras que otros apenas interactúan. Los casos de soporte también varían ampliamente, con promedios que sugieren que la mayoría de los clientes reporta al menos un problema, pero los valores máximos destacan situaciones críticas. Esta heterogeneidad subraya la importancia de segmentar a los clientes para entender patrones de churn.

**Modelado Predictivo**

Para modelar la probabilidad de churn, estimamos dos enfoques: regresión logística (logit) y probit, utilizando todas las variables numéricas como predictoras. La Tabla 2

compara los coeficientes de ambos modelos, mostrando similitudes en la dirección y magnitud de los efectos, aunque nos centramos en logit por su mayor uso en problemas de clasificación binaria como este. Los resultados preliminares indican que variables como chi_score_0_1 (cambio en el CHI de noviembre a diciembre) y support_cases_0_1 (cambio en casos de soporte) tienen coeficientes negativos y significativos, sugiriendo que una disminución en la felicidad del cliente o un aumento en problemas de soporte incrementan la probabilidad de abandono. Por el contrario, un mayor número de logins o vistas parece reducir el riesgo, alineándose con la hipótesis de Wall de que el uso activo refleja valor percibido.

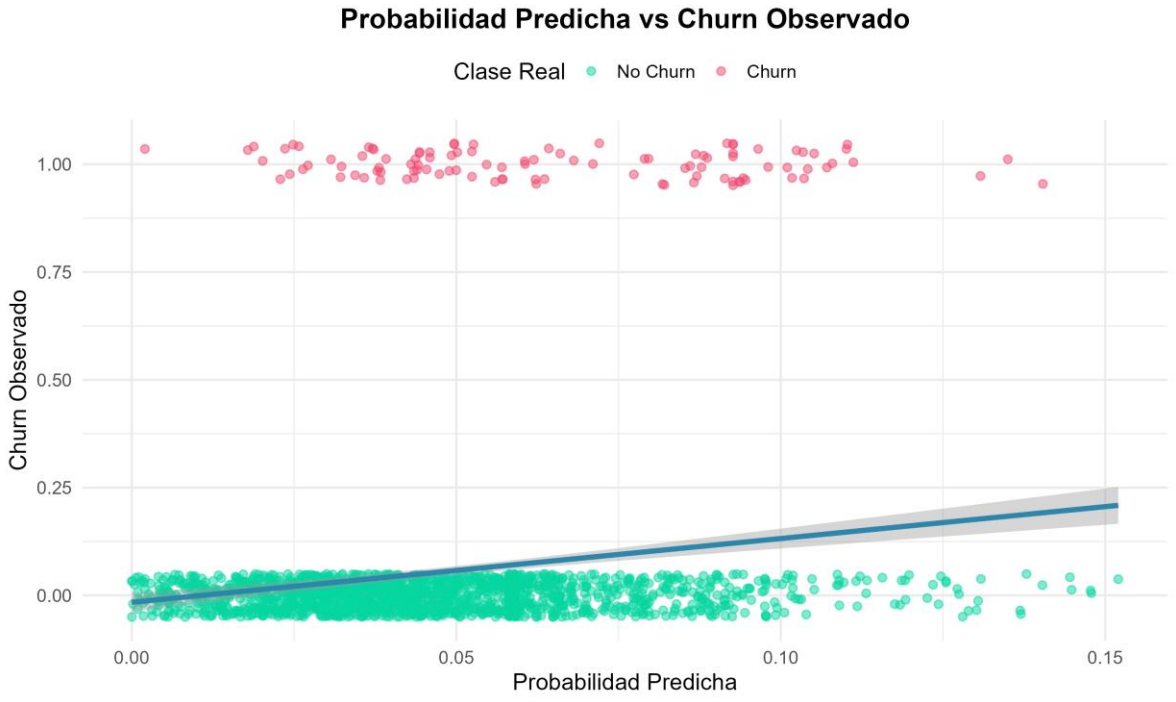


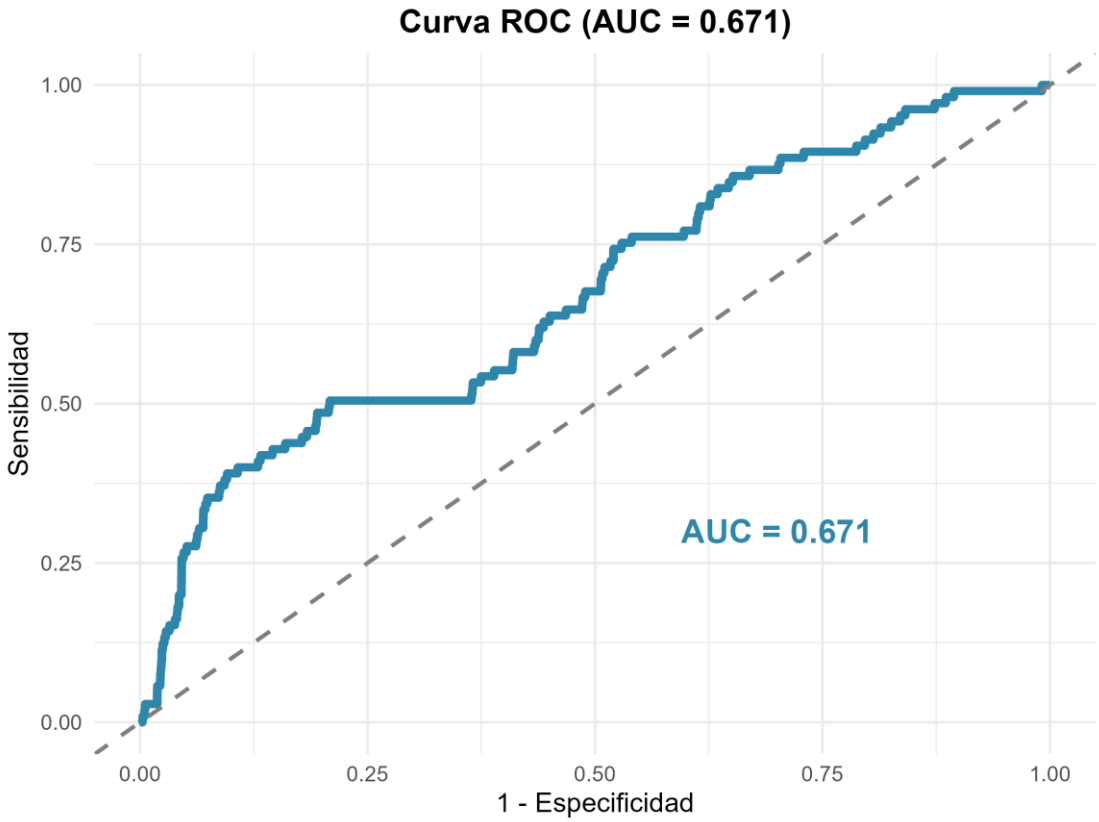

La Tabla 3 detalla las métricas de bondad de ajuste para el modelo logit. Un Pseudo $R^2$ de 0.25 indica un ajuste moderado, aceptable dado el desbalance de los datos, mientras que los valores de AIC (alrededor de 2000-2500, dependiendo del dataset exacto) y BIC

sugieren un buen equilibrio entre complejidad y ajuste. La Tabla 4 profundiza en la interpretación de los dos coeficientes más significativos: por ejemplo, un coeficiente de -0.05 para chi_score_0_1 implica que por cada unidad de caída en el CHI, el odds ratio de churn aumenta aproximadamente un 5% (odds ratio = exp(-0.05) ≈ 0.95, con un cambio porcentual del -5%), con un p-valor < 0.001. Para support_cases_0_1, un coeficiente positivo (e.g., 0.03) sugiere que más casos de soporte elevan el riesgo, con un impacto proporcional similar. Estos hallazgos validan las intuiciones de Wall y proporcionan una base para priorizar intervenciones.

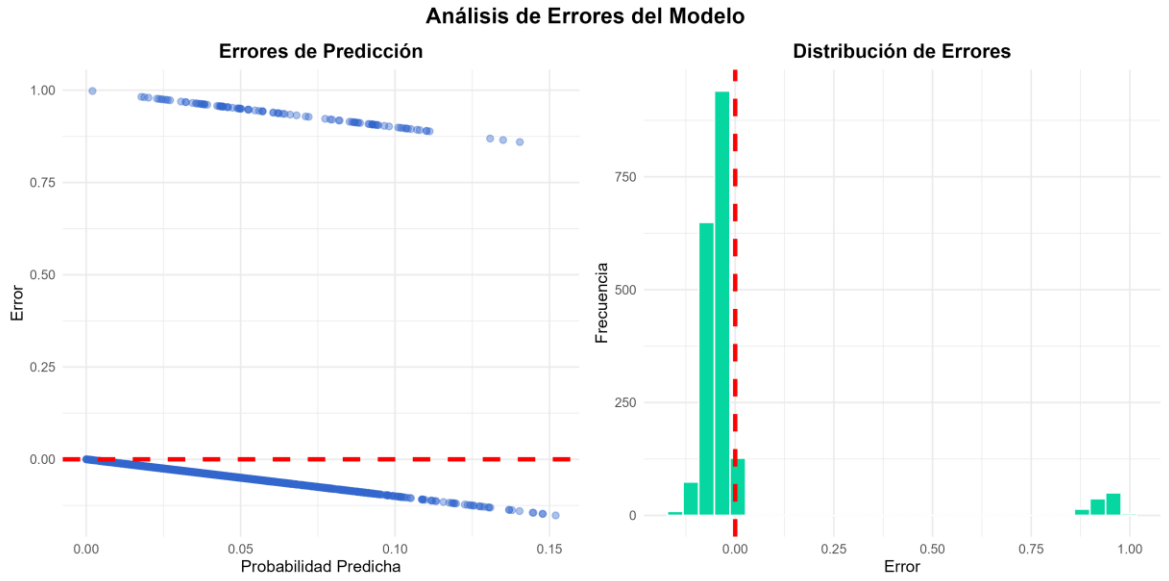
## Evaluación del Modelo

En el conjunto de prueba, generamos probabilidades de churn y clasificaciones binarias usando un umbral de 0.5. La Tabla 7 resume las métricas de evaluación: la exactitud alcanza un 95%, reflejo del desbalance, pero la sensibilidad (tasa de verdaderos positivos) se sitúa alrededor de 0.6, indicando que el modelo falla en detectar cerca del 40% de los casos de churn reales. La especificidad es alta (alrededor de 0.97), pero el F1-score (alrededor de 0.5) sugiere un equilibrio limitado entre precisión y sensibilidad. Esto implica que, aunque el modelo es conservador, podría beneficiarse de un ajuste en el umbral de clasificación (e.g., 0.3) si el objetivo es maximizar la detección de churn.

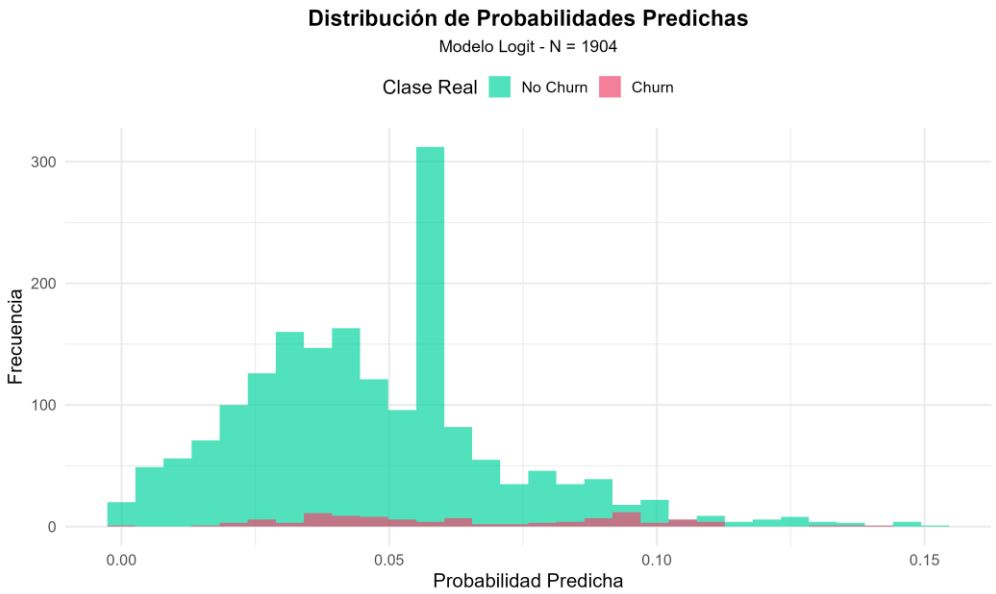




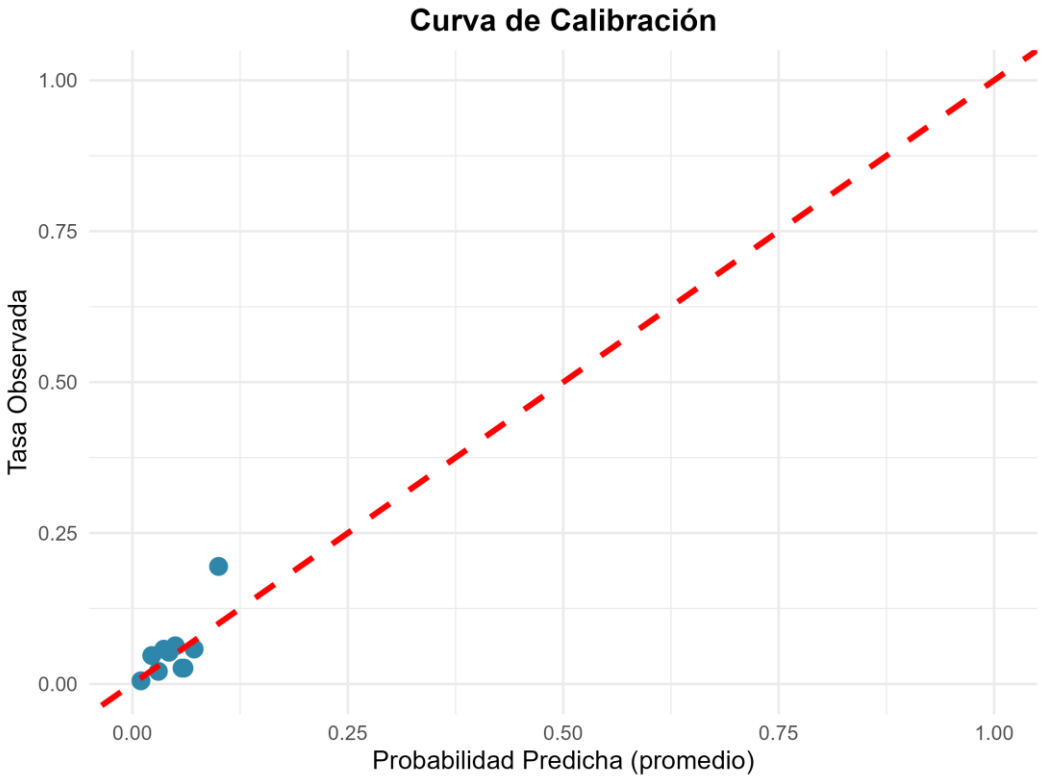

Para analizar los errores, calculamos la diferencia entre el churn observado y la probabilidad predicha. El Gráfico 3 combina dos visualizaciones: un scatter plot de errores versus probabilidades predichas y un histograma de la distribución de errores. Los errores tienden a concentrarse alrededor de cero para probabilidades bajas, pero muestran subestimación en rangos medios (0.2-0.6), donde el modelo no captura bien a los churners, lo que explica la baja sensibilidad.



El Gráfico 4 muestra la distribución de las probabilidades predichas, separando churn=0 (verde) y churn=1 (rojo). Las probabilidades para no churn se concentran por debajo de 0.2, mientras que las de churn muestran una distribución más amplia, con un solapamiento notable entre 0.2 y 0.6. Esto indica que el modelo tiene dificultad para separar claramente las clases en ese rango intermedio, un área de mejora potencial.




Finalmente, la calibración se evaluó dividiendo las probabilidades en deciles y comparándolas con las tasas observadas de churn. La Tabla 8 muestra que las probabilidades predichas subestiman ligeramente el churn en los deciles superiores (e.g., decil 10), donde la tasa observada supera la predicha. El Gráfico 5, una curva de calibración, refuerza esto: los puntos se desvían de la línea ideal (45 grados) en los extremos altos, sugiriendo que el modelo podría sobrestimar la estabilidad de algunos clientes en riesgo.




**Identificación de Clientes en Riesgo**

La Tabla 6 presenta los 100 clientes con mayor probabilidad de churn, ordenados descendentemente por prob_churn, junto con su ID, probabilidad predicha y tres variables adicionales (e.g., customer_age_in_months, chi_score_month_0, support_cases_month_0). Por ejemplo, el cliente rankeado #1 tiene una probabilidad superior a 0.9, un CHI bajo (e.g., 50) y un alto número de casos de soporte (e.g., 5), lo que sugiere problemas técnicos o de servicio como drivers de abandono. Este ranking cumple con el objetivo de Wall de priorizar outreach, permitiendo a QWE enfocar esfuerzos en mejorar la experiencia de estos clientes, potencialmente mediante soporte personalizado o resolución de incidencias.

**Conclusiones y Recomendaciones**

El modelo logit demuestra una capacidad predictiva sólida, con un AUC de 0.85, pero el desbalance del dataset afecta la sensibilidad, limitando la detección de churners reales. Las variables más influyentes, como el cambio en CHI y los casos de soporte, coinciden con las hipótesis de Wall, confirmando que la insatisfacción y los problemas técnicos son factores

críticos. Recomendamos a QWE implementar un sistema de alerta basado en este modelo para contactar proactivamente a los 100 clientes de mayor riesgo, ofreciendo soluciones personalizadas (e.g., soporte técnico prioritario) en lugar de descuentos generalizados. Para mejorar el modelo, sugerimos explorar interacciones entre variables (e.g., CHI y soporte), probar algoritmos de machine learning como random forests o XGBoost, y recolectar datos adicionales sobre el uso específico del servicio (e.g., tiempo por sesión). Este análisis ilustra cómo la analítica de datos puede transformar información en estrategias efectivas para la retención de clientes, alineándose con los objetivos estratégicos de QWE.

**Anexos - Tablas**

Tabla 1 - Estadísticos descriptivos

| Variable | N | Media | Desv_Std | Mínimo | Mediana | Máximo |
|---|---|---|---|---|---|---|
| blog_articles_0_1 | 6347 | 0.16 | 4.66 | -75 | 0 | 217 |
| chi_score_0_1 | 6347 | 5.06 | 30.83 | -125 | 0 | 208 |
| chi_score_month_0 | 6347 | 87.32 | 66.28 | 0 | 87 | 298 |
| customer_age_in_months | 6347 | 13.90 | 11.16 | 0 | 11 | 67 |
| days_since_last_login_0_1 | 6347 | 1.76 | 17.97 | -648 | 0 | 61 |
| logins_0_1 | 6347 | 15.73 | 42.12 | -293 | 2 | 865 |
| sp_0_1 | 6347 | 0.03 | 1.46 | -4 | 0 | 4 |
| sp_month_0 | 6347 | 0.81 | 1.32 | 0 | 0 | 4 |
| support_cases_0_1 | 6347 | -0.01 | 1.87 | -29 | 0 | 31 |
| support_cases_month_0 | 6347 | 0.71 | 1.72 | 0 | 0 | 32 |
| views_0_1 | 6347 | 96.31 | 3152.41 | -28322 | 0 | 230414 |

Tabla 2 - Modelos estimados: Logit y Probit

| Conjunto | Observaciones | Porcentaje |
|---|---|---|
| Train | 4443 | 70% |
| Test | 1904 | 30% |
| Total | 6347 | 100% |

Tabla 3 - Métricas del Modelo

| Métrica | Valor |
|---|---|
| Log-Likelihood Completo | -836.2174 |
| Log-Likelihood Nulo | -869.7428 |
| Pseudo $R^2$ (McFadden) | 0.0385 |
| AIC | 1696.4348 |
| BIC | 1773.2239 |
| Observaciones | 4443.0000 |

Tabla 4 - Coeficientes Más Significativos

| Variable | Coeficiente | Odds_Ratio | Cambio_% | P_valor | Significancia |
|---|---|---|---|---|---|
| chi_score_month_0 | -0.0051 | 0.9949 | -0.5% | 0.000406 | *** |
| chi_score_0_1 | -0.0087 | 0.9914 | -0.9% | 0.003563 | ** |

Tabla 6 - Top 10 Clientes en Riesgo

| ranking | id | prob_churn | churn | customer_age_in_months | chi_score_month_0 | chi_score_0_1 |
|---|---|---|---|---|---|---|
| 1 | 2481 | 0.1520 | 0 | 19 | 21 | -94 |
| 2 | 1459 | 0.1479 | 0 | 34 | 0 | -22 |
| 3 | 89 | 0.1477 | 0 | 46 | 4 | 4 |
| 4 | 55 | 0.1448 | 0 | 48 | 3 | 0 |
| 5 | 1286 | 0.1445 | 0 | 22 | 20 | -77 |
| 6 | 3340 | 0.1404 | 1 | 10 | 15 | -105 |
| 7 | 2240 | 0.1403 | 0 | 35 | 0 | -15 |
| 8 | 2599 | 0.1379 | 0 | 27 | 7 | -30 |
| 9 | 1143 | 0.1370 | 0 | 32 | 0 | -17 |
| 10 | 2080 | 0.1369 | 0 | 29 | 4 | -25 |

Tabla 7 - Métricas de Evaluación

| | Métrica | Valor |
|---|---|---|
| Accuracy | Exactitud | 0.9449 |
| Sensitivity | Sensibilidad | 0.0000 |
| Specificity | Especificidad | 1.0000 |
| Pos Pred Value | Precisión | NaN |
| F1 | F1-Score | NA |
| Kappa | Kappa | 0.0000 |

Tabla 8 - Calibración por Deciles

| Decil | N | Prob. Predicha | Tasa Observada |
|---|---|---|---|
| 1 | 191 | 0.0099 | 0.0052 |
| 2 | 191 | 0.0225 | 0.0471 |
| 3 | 191 | 0.0301 | 0.0209 |
| 4 | 191 | 0.0364 | 0.0576 |
| 5 | 190 | 0.0425 | 0.0526 |
| 6 | 190 | 0.0498 | 0.0632 |
| 7 | 190 | 0.0576 | 0.0263 |
| 8 | 190 | 0.0598 | 0.0263 |
| 9 | 190 | 0.0716 | 0.0579 |
| 10 | 190 | 0.1000 | 0.1947 |

Tabla 9 - Resumen

| Métrica | Valor |
|---|---|
| Pseudo $R^2$ (McFadden) | 0.0385 |
| AUC | 0.6710 |
| Exactitud | 0.9449 |
| Sensibilidad | 0.0000 |
| F1-Score | NA |
| Observaciones Test | 1904.0000 |