Predicting Customer Churn at QWE Inc.

Juan Sebastián Cárdenas Benavides

Paula Ximena Rodriguez Rios

Valery Ramirez Mejia

Universidad Pontificia Javeriana

Analítica de los Negocios

Juan Nicolas Velasquez Rey

Bogotá D.C, Colombia

24 de Octubre de 2025

Este análisis tiene como objetivo evaluar los factores determinantes de la deserción de clientes en QWE Inc., una empresa dedicada a ofrecer servicios de gestión de presencia en línea para pequeñas y medianas compañías. A través de la estimación de un modelo logit de probabilidad de churn, se busca identificar las variables que influyen en la decisión de los clientes de abandonar la plataforma, considerando indicadores clave como el índice de felicidad del cliente (CHI), la actividad de uso del servicio, los casos de soporte y la antigüedad de la relación. Se busca proporcionar una visión integral y basada en datos sobre los factores que explican la pérdida de clientes, con el fin de apoyar a QWE Inc. en el diseño de estrategias preventivas de retención más efectivas y sostenibles en el tiempo.

En la primera fase del análisis se realizó la importación y limpieza de la base de datos. Se seleccionarán librerías como tydiverse, readxl y janitor para garantizar uniformidad. Se seleccionaron las columnas relevantes para el estudio, incluyendo la variable independiente Churn_1_yes_0_no, que indica si el cliente abandonó la empresa en los dos meses posteriores y un conjunto de variables explicativas relacionadas con el comportamiento y la satisfacción del cliente, como customer_age_in_months, chi_score_0_1, entre otras. Y finalmente se construyó un nuevo marco de análisis "model_data", que sirvió como insumo para la estimación del modelo logit.

En la segunda fase se desarrolló un análisis descriptivo con el fin de comprender las características generales de la muestra y las diferencias entre los clientes que desertaron y los que continuaron con el servicio. Para ello, se calcularon medidas estadísticas como medias y desviaciones agrupadas por el estado de churn, utilizando las funciones group_by() y summarise() en R. Los resultados se organizaron en una tabla descriptiva visualizada directamente en el programa mediante la función gt(), la cual permitió comparar los promedios de variables clave entre ambos grupos (Ver Tabla 1).

En promedio, los clientes que abandonaron presentaron un índice de felicidad (CHI) considerablemente menor, una caída reciente en su nivel de satisfacción (ΔCHI negativo), menor actividad en la plataforma (menos logins y visualizaciones) y un mayor tiempo sin ingresar al sistema, en comparación con los clientes retenidos. Estos hallazgos sugieren que la deserción está asociada a la disminución de la satisfacción y al desuso de la plataforma, lo que validó la pertinencia de las variables seleccionadas para la estimación del modelo logit en la siguiente fase.

El modelo logit que estimamos permitió identificar los principales factores que afectan en la deserción de los clientes. En la tabla de las métricas globales (Ver Tabla 2) podemos ver un R^2 de McFadden de 0.0440, lo que significa que el modelo explica el 4.4% de la variabilidad en las decisión de abandono del servicio. Pese al bajo valor comparado con un R^2 de un modelo lineal, esto se debe a que tiene comportamiento binario, especialmente en fenómenos como el churn, donde influyen factores subjetivos (percepción del servicio, competencia, o condiciones del mercado) que no están completamente captados por las variables que tenemos.

En cuanto a la tabla del modelo de regresión (Ver Tabla 3Tabla 3.), el intercepto −2.756, con un $p<0.001$ representa la probabilidad mínima de churn para un cliente promedio sin características observadas de riesgo y sirve como un punto de referencia a partir del cual se evalúan los efectos de las demás variables.

Entre los factores más significativos del modelo, el índice de felicidad actual del cliente (chi_now) presenta un efecto negativo y altamente significativo ($p<0.001$), lo que indica que un mayor nivel de satisfacción disminuye la probabilidad de abandono. Cada punto adicional en chi_now reduce los odds de churn en aproximadamente un 0.5%. De manera similar, el cambio en el índice de felicidad (chi_change) muestra un coeficiente negativo y significativo ($p<0.001$), donde cada punto adicional de este factor reduce la probabilidad de abandono en un 1%, evidenciando que una mejora reciente en la experiencia del cliente reduce aún más la probabilidad de deserción; en contraste, caídas en este indicador constituyen señales tempranas de riesgo. Finalmente, la variable días desde el último ingreso (days_since_last_login_change) presenta un efecto positivo y altamente significativo ($p<0.001$): cada día adicional sin ingresar a la plataforma incrementa los odds de abandono en aproximadamente un 1.7 %, lo que la convierte en un predictor importante de pérdida de compromiso con el servicio.

Por otro lado, se calcularon los efectos marginales promedio, que permiten interpretar los resultados del modelo en términos de probabilidad (ver tabla 4) y no solo en coeficientes estadísticos. Los resultados mostraron que, por ejemplo, cuando aumentan los días sin iniciar sesión (days_since_last_login_change), la probabilidad de que un cliente abandone la plataforma aumenta en aproximadamente 0.0008 puntos porcentuales por unidad. Aunque el valor parece pequeño, en bases de datos grandes como esta, estos cambios acumulados pueden reflejar patrones. Por otro lado, el indicador de experiencia del cliente (chi_change) tuvo un efecto negativo de alrededor de -0.0005, lo que significa que cuando el CHI mejora, la probabilidad de churn disminuye ligeramente. Esto confirma que mejorar la satisfacción del cliente puede ayudar a prevenir su salida. Finalmente, la variable relacionada con cambios en casos de soporte (cases_change), tuvo un efecto positivo de 0.0062, lo que indica que el aumento en casos de soporte puede estar asociado con un mayor riesgo de abandono, posiblemente por experiencias negativas.

Posteriormente, se evaluó qué tan bien el modelo logra predecir quién se va y quién se queda (Ver tabla 5). Para ello, se realizó una comparación entre los valores reales de churn y las predicciones generadas por el modelo, lo cual permitió observar qué tan acertadas eran las decisiones del modelo al clasificar a los clientes. A partir de esta comparación se construyó una matriz de confusión, y con base en ella se calculó la precisión (accuracy), que fue de aproximadamente 94.91%. Esto quiere decir que, de cada 100 clientes, el modelo clasifica correctamente a unos 95. Este resultado muestra que el modelo tiene una muy buena capacidad predictiva general. Sin embargo, también se debe considerar que en los datos hay muchos más clientes que no abandonan que los que se van, lo que puede influir en la alta precisión.

Finalmente, al analizar el gráfico 1 de los residuos estandarizados frente a las probabilidades predichas, se observa que los puntos no se dispersan de manera completamente aleatoria alrededor de la línea cero. Por el contrario, los residuos se agrupan en dos franjas curvas, una ubicada por encima de cero y otra por debajo. Este comportamiento indica que los errores tienden a seguir una forma y no se distribuyen de manera completamente aleatoria. Esa curvatura sugiere que el modelo podría no estar capturando por completo la relación entre las variables independientes y la probabilidad de churn, lo cual podría deberse a la presencia de efectos no lineales o a la falta de algunas variables explicativas. Aunque el modelo presenta un buen nivel de precisión general, este patrón en los residuos evidencia que podrían realizarse ajustes adicionales para mejorar su capacidad de representación del comportamiento real de los clientes.

El modelo logit permitió identificar que la satisfacción del cliente (CHI), su mejora reciente y la actividad en la plataforma son los principales factores que reducen la probabilidad de abandono, mientras que la inactividad, la antigüedad y el aumento en los casos de soporte la incrementan. Aunque el modelo explica solo un 4.4% de la variabilidad, logra una precisión del 94.9%, mostrando un buen desempeño predictivo. Estos resultados ofrecen a QWE Inc. una base sólida para implementar estrategias proactivas de retención, enfocadas en detectar y atender oportunamente a los clientes en riesgo.

Anexos

Tabla 1.

Tabla 1. Estadísticas descriptivas por grupo de churn

Comparación entre clientes retenidos (churn=0) y desertores (churn=1)

Estado del cliente	Número de observaciones	Edad promedio (meses)	CHI promedio	Cambio en CHI	Casos actuales	Prioridad promedio	Cambio en logins	Cambio en views	Cambio días sin login
0	6024	13.81873	88.60591	5.530212	0.7242696	0.8295759	16.13894	106.6096	1.511454
1	323	15.35294	63.27245	-3.736842	0.3715170	0.4995577	8.06192	-95.7678	6.486068

Tabla 2.

Tabla de métricas globales del modelo

Pseudo R^2 de McFadden y medidas de ajuste

Modelo	Pseudo R^2 (McFadden)	AIC	BIC	logLik	n
Modelo Logit	0.0440	2,462.8	2,537.1	−1,220.385	6347

Tabla 3.

Modelo Logit de Probabilidad de Churn

Coeficientes, Odds Ratios e intervalos de confianza

	Coeficiente (sig.)	Std. Error	Odds Ratio (95% CI)	p-value
(Intercept)	-2.756***	0.106	0.064 (0.051 - 0.078)	0.000
customer_age_months	0.012*	0.005	1.013 (1.002 - 1.023)	0.020
chi_now	-0.005***	0.001	0.995 (0.993 - 0.998)	0.000
chi_change	-0.010***	0.002	0.990 (0.985 - 0.994)	0.000
cases_now	-0.116	0.085	0.890 (0.738 - 1.032)	0.174
cases_change	0.132*	0.064	1.141 (1.017 - 1.306)	0.039
priority_now	-0.030	0.074	0.970 (0.839 - 1.123)	0.682
logins_change	0.000	0.002	1.000 (0.996 - 1.004)	0.895
blogs_change	0.001	0.020	1.001 (0.957 - 1.026)	0.980
views_change	-0.000**	0.000	1.000 (1.000 - 1.000)	0.007
days_since_last_login_change	0.017***	0.004	1.017 (1.009 - 1.026)	0.000

Significancia: *** p<0.001, ** p<0.01, * p<0.05

Tabla 4.

Efectos Marginales Promedio del Modelo Logit

Interpretación del impacto promedio de cada variable sobre la probabilidad de churn

factor	AME	SE	z	p	lower	upper
blogs_change	0.0000	0.0009	0.0255	0.9797	-0.0018	0.0018
cases_change	0.0062	0.0030	2.0520	0.0402	0.0003	0.0122
cases_now	-0.0055	0.0041	-1.3568	0.1748	-0.0135	0.0024
chi_change	-0.0005	0.0001	-4.1699	0.0000	-0.0007	-0.0003
chi_now	-0.0002	0.0001	-3.7119	0.0002	-0.0003	-0.0001
customer_age_months	0.0006	0.0003	2.3161	0.0206	0.0001	0.0011
days_since_last_login_change	0.0008	0.0002	3.9584	0.0001	0.0004	0.0012
logins_change	0.0000	0.0001	0.1324	0.8946	-0.0002	0.0002
priority_now	-0.0014	0.0035	-0.4103	0.6816	-0.0083	0.0054
views_change	0.0000	0.0000	-2.6874	0.0072	0.0000	0.0000

Tabla 5.

Matriz de Confusión del Modelo Logit

Comparación entre valores reales y predichos

Real	Predicho	Freq
0	0	6024
1	0	323

Accuracy del modelo: 0.9491

Grafico 1

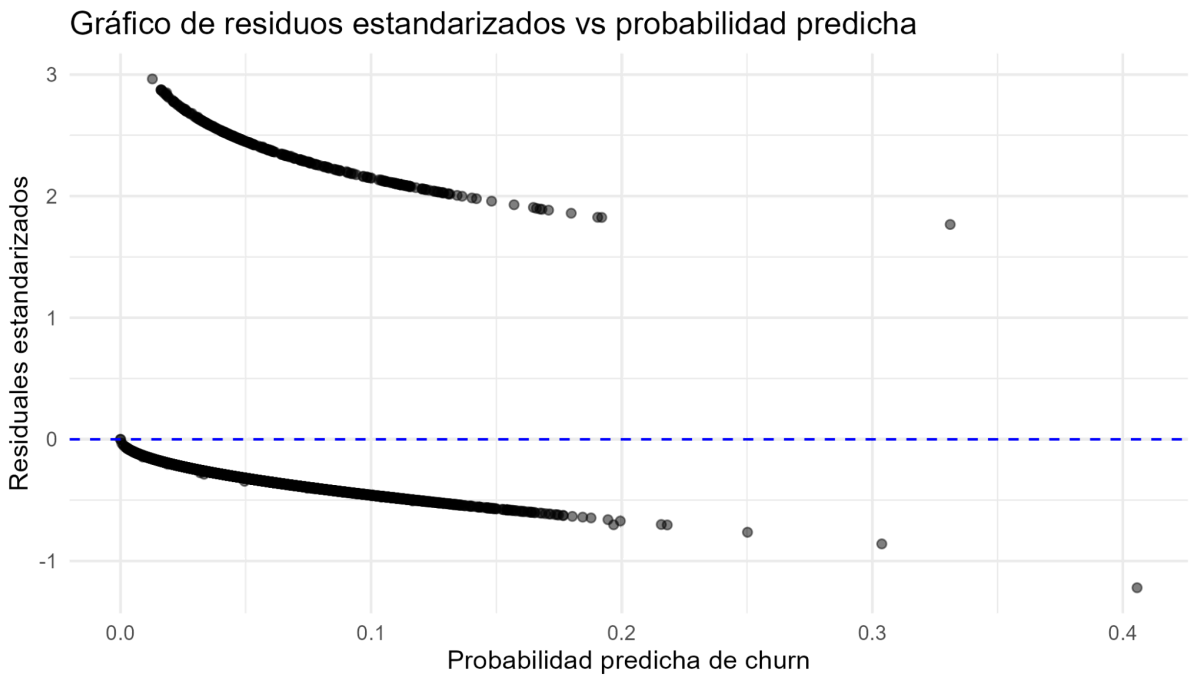
Gráfico de residuos estandarizados vs probabilidad predicha
Residuales estandarizados
3
2
1
0
-1
0.0
0.1
0.2
0.3
0.4
Probabilidad predicha de churn