Predicting Customer Churn at QWE INC.

Julián Mateo Arcos

Felipe Angel

Santiago Velasquez

Profesor: Juan Nicolas Velasquez

Analítica de Datos

Bogotá 2025

Predicting Customer Churn at QWE INC.

Introducción

QWE Inc. es una empresa que ofrece un servicio de suscripción a un mercado objetivo de empresas pequeñas y medianas para gestionar su presencia en internet. La dirección de la empresa ha identificado la necesidad de un análisis sobre la retención de clientes. En la actualidad la empresa reacciona a el abandono de clientes solo cuando estos se comunican para cancelar el contrato, por lo que los gerentes de la empresa necesitan un modelo analítico para estimar la probabilidad de que un cliente abandone la empresa en los próximos dos meses.

La empresa necesita identificar a los clientes con alta probabilidad de fuga y entender los factores que impulsan esa clase de decisiones. El identificar a estos clientes le permitirá a la empresa contactarlos antes de que ellos decidan cancelar el contrato, esto con el fin de evitar el abandono sin tener que incurrir en descuentos en los contratos.

Exploración de los Datos

Para poder lograr la tarea, contamos con una base de datos de más de 6000 clientes de QWE Inc. al 1 de diciembre de 2011. Para el análisis se van a estudiar variables como la antigüedad del cliente, explorando la hipótesis de que aquellos de entre 6 y 14 meses de antigüedad son el grupo más riesgoso. Por otro lado, el Índice de Felicidad del Cliente (CHI), tanto su valor actual como los cambios recientes. Adicionalmente se tendrán en cuenta datos de servicios como el número de casos de soporte. Por último, se revisarán los patrones de uso por medio de las variables de inicio de sesión o visitas. Para iniciar con el estudio se calcularon las estadísticas descriptivas de dichas variables mencionadas, y los resultados fueron los siguientes:

Statistic	N	Mean	St.Dev.	Min	Max
churn	6.347,00	0,05	0,22	0,00	1,00
age_months	6.347,00	13,90	11,16	0,00	67,00
chi_score_m0	6.347,00	87,32	66,28	0,00	298,00
chi_score_change	6.347,00	5,06	30,83	-125,00	208,00
support_cases_m0	6.347,00	0,71	1,72	0,00	32,00
logins_change	6.347,00	15,73	42,12	-293,00	865,00
views_change	6.347,00	96,31	3.152,41	-28322,00	230414,00

Tabla 1. Estadísticas descriptivas Base de Datos QWE Inc.

Modelo Logit

Para poder ayudar a QWE Inc. a reducir los clientes que abandonan su contrato, proponemos un modelo logit el cual en este caso es una herramienta estadística util para predecir la probabilidad de que un cliente en específico vaya a dejar la compañía. El modelo utiliza distintas variables para entender como cada una de ellas impacta la probabilidad de fuga. En nuestro caso las variables a evaluar son:

- La antigüedad del cliente (agrupada en categorías).
- Su índice de felicidad (el CHI Score).
- Su interacción con soporte técnico (casos y prioridad).
- Cambios recientes en su actividad (cambios en logins y views).

Por medio del lenguaje de programación R se realizó un modelo logit que predijera la probabilidad de abandono de los clientes, los resultados se analizaran a continuación.

Interpretación de los Betas

En el caso, la empresa busca anticiparse a la salida de clientes (churn) mediante un modelo logit que relacione variables actuales con la probabilidad de salir dentro de los dos meses siguientes. Los resultados de este modelo reflejan un comportamiento estadístico consistente con lo que planteaba Richard Wall, vicepresidente de servicio al cliente, quien dijo que seguramente los usuarios con entre seis y catorce meses de antigüedad eran el grupo más propenso a abandonar, y que el *Customer Happiness Index* (CHI) debía tener una relación inversa con la deserción.

Los resultados empíricos confirman las ideas que se tenían con anterioridad. Acá, se muestra que la antigüedad del cliente es uno de los factores más determinantes en la probabilidad de cancelación, se evidencio que los clientes con entre 6-14 meses de relación presentan un beta de 2.224 con alta significación (pvalue < 0.01) con un *odds ratio* (OR) de 9.243, lo que indica que un cliente en ese rango de antigüedad tiene nueve veces más probabilidades de desertar que uno con menos de seis meses (grupo base), manteniendo las demás variables constantes. De forma análoga, la categoría “más de 14 meses” tiene un coeficiente 2.000 (OR = 7.386), lo cual deja ver que, aunque la experiencia reduce parcialmente el riesgo, este sigue siendo mayor que en los clientes más nuevos añadiendo que continúa siendo un dato con alta significación dado el pvalue menor a 0.01, lo mencionado sugiere que incluso los clientes más antiguos pueden mostrar señales de insatisfacción con el tiempo.

En cuanto al nivel de satisfacción medido por el CHI, se observa un coeficiente negativo y estadísticamente significativo a todos los niveles ($\beta = -0.011$ y pvalue < 0.01), sugiriendo que, por cada punto adicional en el CHI, la probabilidad de churn se reduce en un 1.1% aproximadamente. No obstante, el cambio reciente en el CHI (0-1), no parece ser significativo, suponiendo que el

valor absoluto de la satisfacción actual explica mejor el modelo que los cambios presentados mes a mes.

Adicionalmente, las variables relacionadas con el servicio técnico y uso del sistema presentan resultados más modestos que los previamente expuestos. En este caso, aparece que el número de casos de soporte y la prioridad asignada a esos casos no cuentan con evidencia estadística significativa, mostrando que los problemas reportados no necesariamente están relacionados con cancelaciones, algo similar ocurre con el número de logins, pues este no tiene un impacto significativo sobre la probabilidad de churn ($\beta = -0.0001$ y pvalue < 0.01.

Para finalizar, el intercepto de -3.833 con pvalue menor a 0.01, refleja que para el grupo de cliente con menos de 6 meses y valores promedio en las demás variables (grupo de referencia), la probabilidad de cancelación es baja como fue previamente expuesto contando con una alta significancia.

Tabla 2: Modelo Logit para Predecir Churn

Dependent variable:	
Probabilidad de Churn (1=Sí)	
Antigüedad (6-14 meses)	**2.224*** **
	(0.224)
Antigüedad (> 14 meses)	**2.000*** **
	(0.231)
CHI Score (Mes 0)	**-0.011*** **
	(0.001)
Cambio CHI (0-1)	**-0.003**
	(0.002)
Casos de Soporte (Mes 0)	**0.0003**
	(0.071)
Prioridad Soporte (SP Mes 0)	**-0.063**
	(0.074)
Cambio Logins (0-1)	**0.001**
	(0.002)
Cambio Vistas (0-1)	**-0.0001*** **
	(0.00004)
Intercepto	*-3.833*** *
	(0.199)
Observations	6,347
Log Likelihood	-1,169.358
Akaike Inf. Crit.	2,356.715
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabla 2. Modelo Logit para predecir Churn.

Matriz de Confusión

Se generaron tres matrices de confusión con umbrales distintos (0.10, 0.15 y 0.20) para examinar cómo se altera el balance entre la sensibilidad (que se refiere a la habilidad de detectar a los clientes que realmente cancelan) y la especificidad (que tiene como objetivo evitar clasificar incorrectamente a los clientes que permanecen).

Umbral de 0.20: Inicialmente, con el umbral de 0.20, el modelo adopta un enfoque bastante conservador, aquí clasifica a la mayoría de los clientes como no propensos al churn. Esto se observa en el número tan elevado de verdaderos negativos de 6022, mientras que el número de verdaderos positivos fue muy bajo, siendo de 3, esto se traduce a que, básicamente no se detectan clientes que realmente abandonan. En términos de negocio, este umbral minimiza las molestias a los clientes leales, pero a costa de no anticipar la salida de casi nadie, pues se trata de un modelo con alta precisión en las predicciones positivas, pero con baja sensibilidad, ya que deja pasar a la gran mayoría de los que efectivamente se van.

		Predicho	
		0	1
Real	0	6022	2
	1	320	3

Tabla 3. Matriz de Confusión Umbral 0.20

Umbral de 0.15: Ahora bien, si el umbral se disminuye a 0.15, el modelo se equilibra y surgen 54 positivos verdaderos y 269 falsos positivos. Esto indica que en esta situación el modelo es más efectivo detectando a los clientes en riesgo, pero sin perder un control razonable sobre las falsas alarmas. En otras palabras, este punto medio permite una actuación proactiva sobre un conjunto importante de clientes en riesgo sin incomodar innecesariamente a los usuarios satisfechos. Esto significa que se alcanza un equilibrio razonable entre sensibilidad y especificidad, lo que convierte esta decisión en una de carácter estratégico más que netamente estadístico.

		Predicho	
		0	1
Real	0	5808	216
	1	269	54

Tabla 4. Matriz de Confusión Umbral 0.15

Umbral de 0.10: Finalmente, se implementó un umbral de 0.10, donde el modelo cuenta con un comportamiento más agresivo, dado que el número de verdaderos positivos aumenta

considerablemente, significando que ahora se detecta a un mayor porcentaje de clientes que efectivamente cancelan, pero este incremento no es únicamente positivo, pues trae consigo un fuerte aumento en la cantidad de falsos positivos, donde en práctica, trae la implicación de que se molestaría a una mayor cantidad de clientes que no pensaban cancelar.

		Predicho	
		0	1
Real	0	5262	762
	1	188	135

Tabla 5. Matriz de Confusión 0.1

En síntesis, el umbral de 0.20 supone un contexto de "no molestar a nadie" pero con una elevada omisión de los casos verdaderos de churn; por su parte, el umbral de 0.10 es un enfoque "incomodante pero efectivo" que maximiza la detección a expensas de saturar al equipo de retención; finalmente, el umbral de 0.15 se establece como el punto óptimo, en el cual el modelo consigue identificar una proporción significativa de clientes que efectivamente se marcharán, sin provocar un exceso de intervenciones. Esta opción del umbral intermedio tiene una justificación tanto estadística como desde el punto de vista operativo y de la experiencia del cliente.

Distribución de Probabilidades Predichas

Para evaluar la capacidad predictiva decidimos graficar la distribución de las probabilidades predichas. En el grafico la curva roja representa a los clientes que no abandonaron. En este caso Podemos observar un buen resultado ya que vemos que la mayoría de la curva esta agrupada en el extremo Izquierdo, con probabilidades de fuga o churn muy cercanas a 0, confirmándonos que el modelo logit los identifica correctamente como clientes seguros.

Por otro lado, la curva azul que representa a los clientes que, si abandonaron esta mucho más extendida a la derecha, lo que nos confirma que el modelo si tiene cierto poder predictive ya que asigna en promedio probabilidades más altas a los clientes que se fueron en comparación a los que se quedaron.

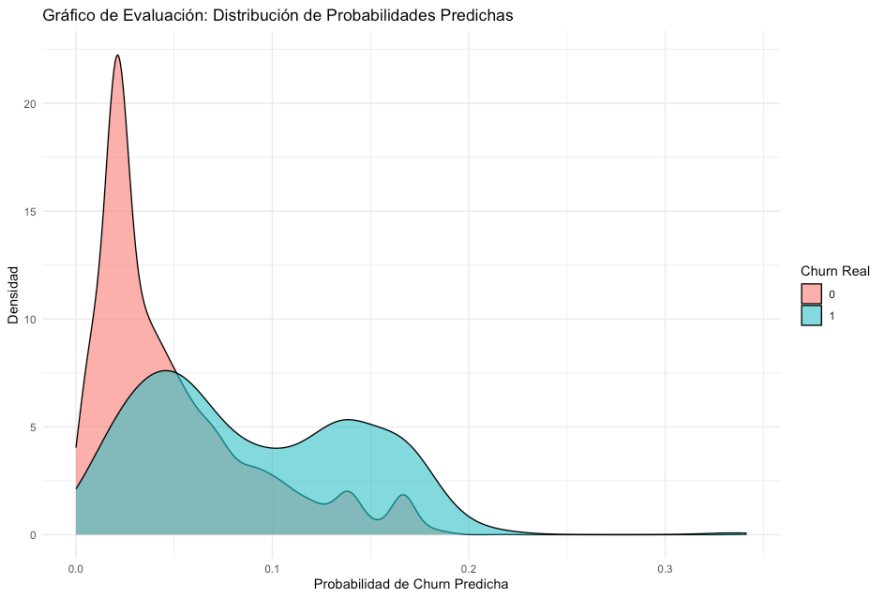

Gráfico 1. Distribución de Probabilidades predichas.

Tras revisar y analizar el grafico llegamos a la conclusión de que, aunque el modelo demuestra tener una capacidad aceptable de predicción, todavía hay una gran porción de la curva azul que todavía se encuentra en la zona de probabilidades baja, camuflándose con la mayoría de la curva roja que representa a los clientes que no abandonaron. Esto es una representación grafica de lo antes encontrado con las matrices de confusión.

Evaluación del Modelo

Por último, decidimos graficar la curva ROC para evaluar la capacidad de predicción del modelo propuesto. Al realizar esto encontramos que el área bajo la curva ROC es de 0,732. Teniendo en cuenta que un AUC por encima de 0,7 indica una discriminación aceptable entre clientes propensos y no propensos a desertar, lo que valida su utilidad para segmentar la base y priorizar acciones de retención. Teniendo eso claro podemos interpretar que si tomamos un cliente al azar que si abandono y un cliente al azar que no abandono, hay un 73,2% de probabilidad de que el modelo haya asignado una puntuación de riesgo más alta al cliente que si se fue.

Adicionalmente se calculo el Pseudo R2 de Mc Fadden, el cual dio como resultado 0,084. Esto es un resultado un cuanto débil ya que nos dice que nuestro modelo solo logra explicar el 8,4% de la varianza total en la decisión de los clientes en hacer churn. Sin embargo, teniendo en cuenta que para los modelos logit el R2 es mucho más bajo que para los modelos tradicionales, es un R2 aceptable para un entorno académico.

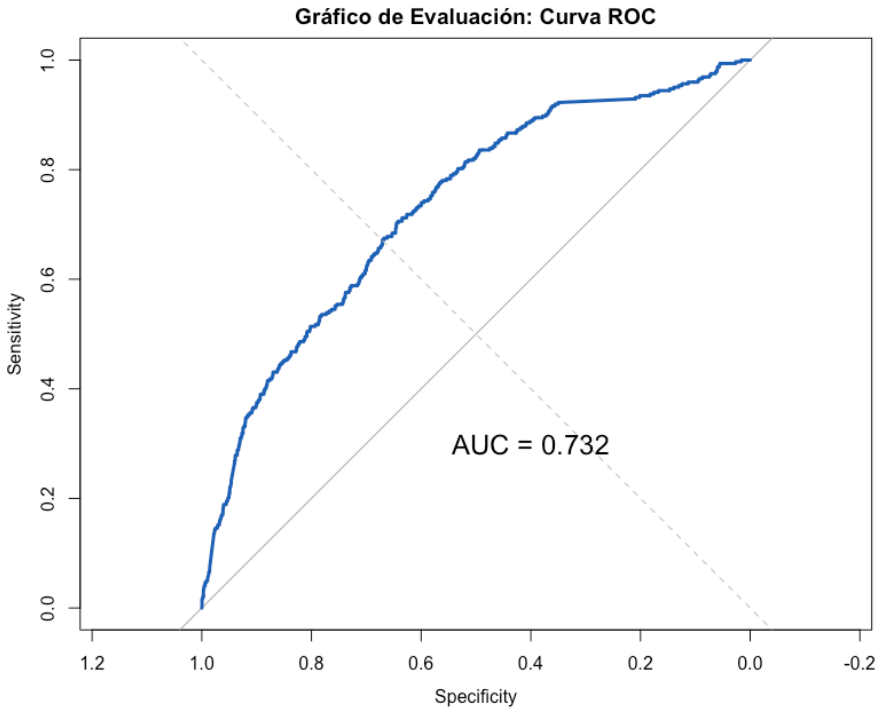

Gráfico 2. Curva ROC

La matriz de confusión con un umbral de 0.15 muestra una exactitud del 92.36 %, aunque el elevado desbalance de clases explica que la precisión sea alta y la sensibilidad más limitada. Cuando el umbral se reduce a 0.10, la sensibilidad mejora a costa de una ligera caída en la precisión, confirmando que el modelo puede ajustarse estratégicamente según los costos relativos de falsos positivos y falsos negativos.

Conclusiones

El modelo logit desarrollado proporciona QWE Inc. Una herramienta para cambiar su estrategia de retención de una reactiva a una proactiva. El análisis cuantitativo valido algunas predicciones de la gerencia y les va a permitir generar una ruta de acción. Algunos de los descubrimientos más importantes son los siguientes. El modelo confirmo que la antigüedad es un predictor esencial. Los clientes en un rango de 6 a 14 meses tienen una probabilidad de abandono significativamente mayor que clientes nuevos. Adicionalmente se encontró que con el puntaje CHI también se puede predecir de buena manera, a mayor felicidad menos probabilidad de fuga.

El modelo demostró tener un rendimiento aceptable, esto se confirma con el área bajo la curva que registra un valor aceptable de 0,732. Sin embargo, a pesar de demostrar ser un buen predictor, tiene un Pseudo R2 de 0,084 que es mediocre. Sin embargo, consideramos que el uso estratégico del umbral de predicción es la clave para ofrecer una buena herramienta de predicción para la

compañía. Nosotros recomendamos el uso de 0,15 como umbral, pero la compañía puede ajustarlo para diferentes fines.