HOW TO
CHARM YOUR CHURN

Predicción y prevención del churn en clientes de ecommerce basado en Datos de Olist Brazil

Nerea López Ziluaga
Data Science Full Time - The Bridge Bilbao

1. Introducción

El churn, o tasa de abandono de clientes, es un indicador clave en el sector ecommerce. Conocer qué clientes corren riesgo de dejar de comprar, predecir el churn, permite a las empresas actuar proactivamente creando estrategias para fidelizar, personalizar ofertas y reducir así la pérdida de ingresos optimizando el ROI.

Estudios del sector muestran que adquirir un nuevo cliente puede costar entre 5 y 7 veces más que retener uno existente. En el contexto español, datos del informe de eCommerce de IAB Spain y del Observatorio Nacional de Tecnología (ONTSI) señalan que el coste medio por adquisición (CPA) puede alcanzar los 35-50€ por cliente, mientras que estrategias de retención pueden reducir ese coste hasta un 70%.

En este contexto, **diseñar un modelo que identifique a los clientes con mayor probabilidad de abandono resulta clave para la eficiencia comercial y el crecimiento sostenible** de un ecommerce.

2. Información sobre Olist Brazil

El **conjunto de datos utilizado en este análisis proviene de Olist y contiene información real de más de 100.000 pedidos históricos**. Estos pedidos incluyen detalles sobre clientes, productos, pagos, envíos y valoraciones.

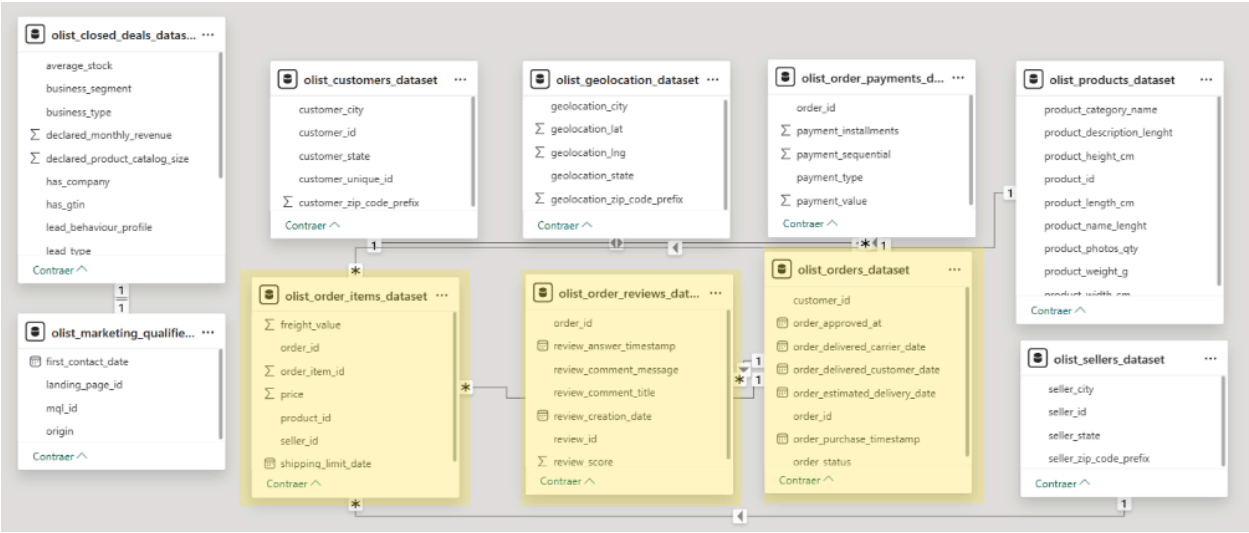
Olist es una empresa brasileña que actúa como integradora de marketplaces. Su plataforma permite a pequeñas y medianas empresas vender productos en grandes marketplaces como Amazon. Olist proporciona herramientas de gestión, pagos, logística y atención al cliente, centralizando toda la experiencia del ecommerce en un solo lugar.

3. Descripción breve de los datasets

El conjunto de datos proporcionado por Olist está compuesto por **varios archivos CSV interrelacionados**, que en conjunto permiten reconstruir el recorrido completo de un pedido: desde que se realiza la compra hasta que se entrega y es evaluada por el cliente.

Dataset	Descripción

olist_orders_dataset.csv	Información de los pedidos (fechas, estado, ID del cliente).
olist_order_items_dataset.csv	Detalles de los productos incluidos en cada pedido.
olist_order_reviews_dataset.csv	Valoraciones realizadas por los clientes.
olist_order_payments_dataset.csv	Métodos de pago y cantidades.
olist_customers_dataset.csv	Información básica de los clientes.
olist_products_dataset.csv	Información sobre los productos.
olist_sellers_dataset.csv	Información sobre los vendedores.
olist_geolocation_dataset.csv	Coordenadas geográficas de códigos postales.
product_category_name_translation.csv	Traducción de categorías de producto del portugués al inglés.



Datasets utilizados en este análisis

Para el análisis de churn, se seleccionan los datasets más directamente relacionados con **el comportamiento del cliente, la experiencia de compra y la satisfacción**. Concretamente:

- **`olist_orders_dataset.csv`**
  Es el eje principal del análisis. Contiene la relación entre cliente y pedido, así como fechas clave (compra, entrega, estimado). Dataset utilizado para definir el churn en función de la última fecha de compra.

- **`olist_order_items_dataset.csv`**
  Utilizado para calcular el **valor de cada pedido** (sumando el precio de los productos). A partir de ahí, se obtiene el **valor medio de compra por cliente**, una variable clave en el churn score.

- **`olist_order_reviews_dataset.csv`**
  Permite medir el **grado de satisfacción** de cada cliente a través del `review_score`, variable también clave para anticipar abandono con el churn score.

De estos, se seleccionaron y transformaron solo las variables necesarias para el análisis, priorizando aquellas relacionadas con el comportamiento del cliente y su experiencia post-compra.

El dataset final resultante contiene una fila por cliente, e incluye:

- Si está churned o no.
- Variables escaladas de valor medio de compra, tiempo de envío y reseña.
- El churn score calculado.
- El segmento de riesgo asignado.

4. Objetivo del EDA

El objetivo principal del análisis exploratorio (EDA) es **identificar patrones y factores asociados al abandono de clientes (churn)** y **construir un modelo de puntuación (churn score)** que permita clasificar a los clientes según su nivel de riesgo. A partir de ahí, se podrán diseñar estrategias de retención específicas para cada tipo de cliente.

5. Hipótesis a confirmar

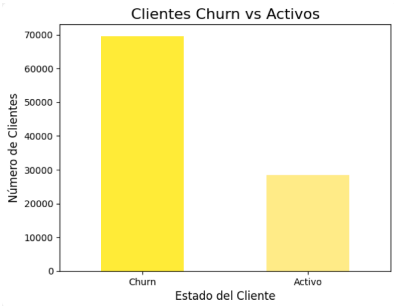
Se plantean las siguientes hipótesis, que se contrastarán mediante análisis estadístico y visual:

Variable	Hipótesis	Relación esperada
Valor medio de compra (avg_order_value)	Los clientes que gastan menos tienen más probabilidad de abandonar	- gasto medio = + churn
Tiempo de entrega (delivery_time)	Los clientes con entregas más lentas tienen más probabilidad de abandonar	+ demora = + churn
Puntuación media de reseña (review_score)	Los clientes que dan puntuaciones bajas tienen más probabilidad de abandonar	- score = + churn

6. Proceso de trabajo

6.1. Definición del churn

Se ha definido como cliente "churned" aquel que no ha realizado ninguna compra en los últimos 180 días desde la última fecha del dataset. Esta métrica se usa como **variable objetivo binaria** (churn).



6.2. Selección de variables y confirmación de hipótesis

Se identifican tres dimensiones directamente relacionadas con la experiencia de cliente:

- Valor medio del ticket de compra
- Tiempo de envío
- Puntuación media de reseñas

Se ha segmentado la muestra entre **clientes activos** (churn = 0) y **clientes churned** (churn = 1), y se han realizado **tests estadísticos (t-test de muestras independientes)** para verificar si existen diferencias significativas entre ambos grupos en tres variables clave:

Valor promedio de compra

Resultado del t-test:
t = -2.456, p = 0.0140

Interpretación:
Existe una diferencia estadísticamente significativa entre los dos grupos. Los clientes que han abandonado tienden a gastar menos por compra que los clientes activos. Esto sugiere que un menor nivel de gasto puede estar relacionado con una menor fidelización.

Tiempo de entrega

- **Resultado del t-test**:
  t = 75.180, p = 0.0000

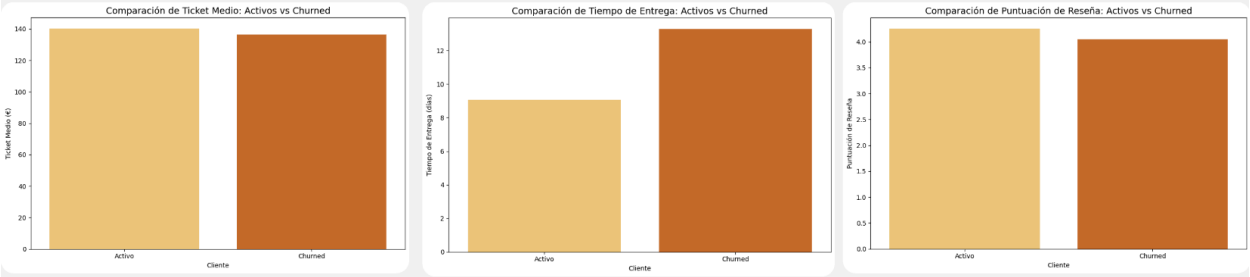
- **Interpretación**:
  Esta diferencia es altamente significativa. Los clientes churned experimentaron tiempos de entrega considerablemente mayores. Esto apoya la hipótesis de que las demoras logísticas pueden ser un factor clave en la pérdida de clientes.

Puntuación de reseña

- **Resultado del t-test**:
  t = -23.016, p = 0.0000

- **Interpretación**:
  También se confirma una diferencia muy significativa. Los clientes churned han otorgado puntuaciones más bajas, indicando una experiencia de compra menos satisfactoria. La satisfacción del cliente está directamente relacionada con la probabilidad de retención.

En conclusión, los resultados apoyan las hipótesis planteadas: **los clientes que abandonan presentan peor comportamiento en todas las variables analizadas**, lo cual refuerza la validez del modelo de churn score y permite justificar las acciones que se propongan en la estrategia final.



6.3. Escalado de variables

Para poder combinarlas en un único índice, las variables se normalizan utilizando MinMaxScaler, llevándolas a una escala de 0 a 1.

6.4. Creación del churn score

Se ha construido un **modelo de scoring** que combina las variables (review_score, delivery_time y avg_order_value) en una escala normalizada (0 a 1), asignando mayor peso al tiempo desde la última compra. La fórmula combina las tres variables disponibles escaladas mediante MinMaxScaler, dando lugar a una métrica continua: **churn_score**.

6.5. Definición de 3 segmentos de riesgo

En base al **churn score**, se divide a los clientes en tres segmentos utilizando los cuartiles (percentiles 25 y 75) como referencia, segmentando a los clientes en tres grupos:

🟢 **Bajo Riesgo**: clientes con un churn score inferior al primer cuartil (Q1). Son clientes fidelizados y con buena experiencia.

🟡 **Riesgo Medio**: clientes entre el primer cuartil (Q1) y el tercero (Q3). Tienen señales mixtas y podrían escalar a riesgo alto si no se interviene.

🔴 **Alto Riesgo**: clientes con un churn score superior al tercer cuartil (Q3). Son los más propensos a abandonar y requieren atención prioritaria.

	customer_id	risk_segment	churn_score	churn
0	00012a2ce6f8dcda20d059ce984917(	Alto riesgo	0.72	1
1	000161a058600d5901f007fab4c2714	Alto riesgo	0.49	1
2	0001fd6190edaaf884bcaf3d49edf07!	Bajo riesgo	0.4	1
3	0002414f95344307404f0ace7a26f1d!	Riesgo medio	0.44	1
4	000379cdec625522490c315e70c7a9f	Alto riesgo	0.49	1
5	0004164d20a9e969af783496f340865	Alto riesgo	0.71	1
6	000419c5494106c306a97b56357480	Alto riesgo	0.76	1
7	00046a560d407e99b969756e0b10f2	Bajo riesgo	0.41	1
8	00050bf6e01e69d5c0fd612f1bcfb69(	Riesgo medio	0.42	1
9	000598caf2ef4117407665ac3327513	Bajo riesgo	0.38	0

6.6. Próximo paso: Estrategia de retención

Sobre la base de esta segmentación, se diseñará una estrategia personalizada para cada segmento, con propuestas de acciones para retener a los clientes más valiosos, reactivar a los

7. Conclusiones

- El churn es un **problema real y medible** en Olist: un porcentaje relevante de clientes ha dejado de comprar tras su primera compra o en los últimos 6 meses.
- Las variables clave que influyen en el churn son claras y han sido **confirmadas mediante test estadístico t-test**:
- Menor valor promedio de compra.
- Mayor tiempo de entrega..
- Peores puntuaciones de reseñas
- Se ha creado un churn score personalizado que permite identificar qué clientes están en mayor riesgo de abandono.
- Se han segmentado los clientes en 3 niveles de riesgo (bajo, medio y alto) según su churn score, lo que permite personalizar estrategias de retención.
- Tenemos una base sólida para tomar decisiones: este análisis permite reducir costes de adquisición actuando sobre los clientes que ya tenemos, con estrategias dirigidas y eficaces.