Autor: Dimitrio Mandamadiotis Kalfagiannis CI 23723158

Tutor: Yaneth Moreno

**DESARROLLO DE UN SISTEMA PARA INFERIR LA DESERCIÓN DE LOS CLIENTES EN UN E-COMMERCE**

**CAPITULO 2**

En éste capítulo, se pretende explicar las bases teóricas necesarias para la comprensión del diseño e implementación de un sistema para inferir la deserción de clientes en un e-commerce.

**2.1 E-Commerce**

Se puede definir al comercio electrónico o E-Commerce como transacciones comerciales realizadas digitalmente entre individuos y organizaciones. Las transacciones comerciales implican un intercambio de valor (por lo general dinero) o comercio entre las partes interesadas a cambio de productos o servicios. El hecho de que se realicen las transacciones digitalmente supone el uso de tecnologías digitales como la internet, la web, dispositivos electrónicos y software.

**2.1.1 Cliente**

Un cliente en el contexto del comercio electrónico es un individuo que realiza compras regulares a través de algún canal digital de una organización. El proceso de compra de este cliente es similar al de una tienda física y consiste en: visualizar los productos o servicios que ofrece la organización, luego procede a agregar lo que desea comprar a un carrito de compras, cuando el cliente está listo para pagar, el carrito es verificado y luego pagado de forma digital, finalmente se genera un registro de pedido en el sistema el cual es procesado por la organización para satisfacer la compra.

**2.1.2 Pedido**

Un pedido u orden de compra es un registro que contiene información relevante sobre una transacción de compra realizada por un cliente en una plataforma de comercio electrónico de una organización. Este registro suele contener información relevante a los procesos de cada organización pero los datos indispensables que incluye son: la fecha en que se realizó un pedido, el valor monetario del mismo y el cliente que lo generó.

**2.1.3 Negocio Contractual**

Se considera negocio contractual a toda organización que interactué y trabaje con sus clientes bajo el modelo de suscripción u otra configuración contractual. En este tipo de organizaciones es relativamente fácil determinar si un cliente está activo o no, ya que por lo general se registran los datos de vencimiento y renovación de contratos o suscripciones. Algunos ejemplos de negocios contractuales son las compañías de seguros, empresas de telecomunicaciones, gimnasios y plataformas SaaS.

**2.1.4 Negocio No Contractual**

Un negocio no contractual es cuando la organización no tiene mecanismo que le ayude a determinar cuando la relación con un individuo o cliente ha finalizado, esto debido a que un cliente puede estar en un periodo de espera antes de realizar otra compra o puede que no compre de nuevo a la organización. Ejemplos de este tipo de negocio son las tiendas de comercio electrónico y las tiendas de venta al por menor.

**2.1.5 Deserción de Clientes**

Deserción de clientes o Churn es la pérdida de clientes debido al desgaste y es lo opuesto a la retención de clientes. Ésta se considera una métrica crucial para las empresas y se utiliza como base para calcular otras métricas de gran importancia como la vida útil del cliente, el valor de vida útil (CLV por sus siglas en inglés), el gasto de adquisición de clientes y el gasto de éxito del cliente.

**2.1.6 Valor de vida útil de un cliente (CLV)**

Esta medida representa el valor que tiene un cliente para un negocio a lo largo de toda la relación, es decir, es el valor total que ha aportado un cliente al negocio desde que realizó una primera compra, hasta que desertó.

**2.2 Modelo**

En líneas generales, un modelo es una representación de un sistema, éste se compone por conceptos que ayudan a explicar el sistema, estudiar los efectos de diferentes componentes y hacer predicciones sobre el comportamiento del mismo. Estos conceptos suelen manifestarse en un grupo de ecuaciones matemáticas

**2.2.1 Modelo Estadístico**

Un modelo estadístico es un modelo matemático que representa una o varias suposiciones estadísticas, estas suposiciones tienen describen propiedades que nos permiten calcular la probabilidad de que ocurra un evento.

**2.2.2 Modelos Buy-Till-You-Die (BTYD)**

Son una familia de modelos estadísticos que tienen como objetivo describir el comportamiento de compra de los clientes como una serie de distribuciones. Estos modelos se basan en dos suposiciones básicas:

1. Los clientes interactúan con la organización cuando están activos o "vivos".
2. Los clientes pueden “morir” o transformarse en clientes inactivos después de un periodo de tiempo.

Existen varios modelos BTYD, pero todos tienen las siguientes características:

**2.2.2.1 Modelado Probabilístico**

Todos estos modelos generan distribuciones de probabilidad para describir el comportamiento de los clientes. Si bien los modelos difieren en cómo crean estas distribuciones o cómo se ven estas distribuciones, todos comparten este objetivo principal.

**2.2.2.2 Tabla de pedidos**

La entrada requerida para generar las predicciones en todos los modelos BTYD es simplemente una tabla que contenga los pedidos de los clientes. Cada modelo procesa esta tabla de manera diferente, pero todos comparten una estructura de entrada común. Debido a este requisito de entrada, cuantos más datos longitudinales se tenga sobre los pedidos de los clientes, más sólidas serán las predicciones del modelo.

**2.2.2.3 Análisis RFM**

El modelado y análisis de RFM es una estrategia común empleada actualmente por las empresas para capturar el comportamiento del cliente y clasificarlo según tres variables:

* **Recencia (Recency):** ¿Hace cuánto tiempo compró un cliente? Este dato hace referencia a la duración en periodos de tiempo entre la primera compra del cliente y la última.
* **Frecuencia (Frecuency):** ¿Con qué frecuencia/consistencia compra un cliente? Este valor representa la cantidad de periodos de tiempo en los que el cliente realizó un pedido.
* **Valor Monetario (Monetary):** ¿Cuánto gasta un cliente en promedio? Esto es la suma de todos los valores monetarios de las compras de un cliente dividida por el número total de compras.

Cada uno de estos elementos del análisis RFM juega un papel clave en las distribuciones del los modelos BTYD.

**2.2.3 Modelo Pareto / NBD**

El nombre se refiere a la distribución de Pareto utilizada para modelar la deserción de clientes y la distribución binomial negativa (NBD) utilizada en la predicción de compras futuras. Este modelos asume que los clientes compran a un ritmo constante durante un periodo de tiempo y luego pasan a la inactividad.

El objetivo del modelo Pareto / NBD es desarrollar el análisis RFM capturando el contexto del cliente al decidir si, el cliente ha desertado o no y con qué frecuencia comprará. Por ejemplo, si estudiamos un cliente A y un cliente B:

* El cliente A compró en una tienda de comercio electrónico cada dos días durante 2 meses. Sin embargo, no ha comprado en las últimas 2 semanas.
* El cliente B compró en la tienda aproximadamente una vez al mes durante los últimos 6 meses. Tampoco ha comprado en las últimas 2 semanas.

A pesar de que ambos clientes tienen la misma antigüedad (14 días), la diferencia en el comportamiento de compra cambia la probabilidad de deserción. Además, a pesar de comprar en la misma tienda, exhiben patrones de compra significativamente diferentes (cada dos días frente a cada mes), por lo que cualquier predicción debe tener en cuenta sus hábitos de compra.

El modelo de Pareto/NBD se basa en cinco supuestos:

Desde el punto de vista del cliente individual:

1. **Compras siguen la distribución Poisson:** Mientras está activo, el número de transacciones realizadas por un cliente en un período de tiempo de longitud *t* se representan mediante una distribución Poisson con media *λt*. En realidad, Pareto/NBD, como su nombre lo indica, utiliza una distribución binomial negativa de la cual, la distribución de Poisson es un caso especial. La distribución de Poisson es una elección acertada para modelar el tiempo entre transacciones, ya que diferentes negocios tendrán diferentes tiempos y los clientes también pueden exhibir diferentes patrones de compra.
2. **“Vida útil” exponencial:** Cada cliente tiene una "vida útil" no observada de longitud *τ*. Este punto en el que el cliente se vuelve inactivo se distribuye exponencialmente con una tasa de deserción *μ*. Esto lo que nos dice es que en general, la mayoría de los clientes se transformarán en inactivos relativamente rápido después de realizar una compra inicial pero una pequeña cohorte de clientes permanecerá durante mucho tiempo.

Heterogeneidad entre clientes:

1. **Las tasas de compra individuales siguen la distribución gamma:** la tasa de compra *λ* para los diferentes clientes sigue una distribución gamma entre la población de clientes.
2. **Tasa de deserción sigue distribuciones gamma diferentes:** Las tasas de deserción *μ* siguen una distribución gamma que es diferente entre los clientes.
3. **Independencia de tasas:** Las tasas de transacciones *λ* y las tasas de deserciones *μ* están distribuidas independientemente entre ellas.

Este modelo tiene mucho poder a la hora de analizar el comportamiento de clientes y requiere solo dos entradas sobre cada cliente, su "recencia" y "frecuencia". El modelo utiliza la inferencia de probabilidad logarítmica para hacer predicciones sobre el futuro. Un problema de este modelo es que su implementación puede ser desafiante desde el punto de vista computacional debido al uso prolongado de la función hipergeométrica gaussiana.

**2.2.4 Modelo BG / NBD**

Este modelo deriva del modelo Pareto / NBD como una alternativa simple de calcular desde el punto de vista computacional. Su nombre hace referencia a la distribución geométrica beta utilizada para modelar la deserción de clientes y la distribución binomial negativa (NBD) utilizada en la predicción de compras futuras.

La mayoría de los aspectos del modelo BG / NBD reflejan directamente los del modelo Pareto / NBD. La única diferencia radica en las consideraciones que definen cómo y cuándo los clientes se vuelven inactivos. La distribución de tiempo de Pareto asume que la deserción puede ocurrir en cualquier momento, independientemente de la ocurrencia de compras reales, en este modelo se asume que el abandono ocurre inmediatamente después de una compra, permitiendo así modelar este proceso mediante la distribución geométrica beta (BG por sus siglas en inglés).

El modelo de BG / NBD se basa en cinco supuestos :

1. Mientras está activo, el número de transacciones realizadas por un cliente representan mediante una distribución Poisson con una tasa de transacciones *λ*. Esto es equivalente a suponer que el tiempo entre transacciones se distribuye exponencialmente con tasa de transacción λ.
2. La tasa de compra *λ* para los diferentes clientes sigue una distribución gamma entre la población de clientes.
3. Después de cualquier transacción, un cliente se vuelve inactivo con probabilidad *p*. Por lo tanto, el punto en el que el cliente "abandona" se distribuye entre las transacciones de acuerdo con una distribución geométrica (desplazada).
4. Las tasas de transacciones *λ* y las tasas de deserciones *μ* están distribuidas independientemente entre ellas.
5. La heterogeneidad en *p* sigue una distribución beta.

El modelo BG / NBD se puede implementar de una manera relativamente sencilla y eficiente, además de que la estimación de sus parámetros no requiere ningún software especializado ni la evaluación de funciones matemáticas no convencionales.

**2.3 Data**

Datos o data en inglés, es una representación simbólica de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades, es decir, transmiten información.

Los datos son indispensables en la investigación científica, las finanzas y en prácticamente cualquier otra forma de actividad organizacional humana. Ejemplos de conjuntos de datos incluyen precios de acciones, tasas de delincuencia, pedidos de clientes en un e-commerce, tasas de desempleo, tasas de alfabetización y datos del censo.

**2.3.1 Dataset**

Un Dataset se puede definir como un conjunto de información en particular, representado en una tabla o matriz de análisis. La tabla está conformada por columnas y cada una de ellas representa una variable de datas, las filas suponen un grupo de datos específicos.

## **2.4 Herramientas Tecnológicas**

**2.4.1 Python**

Python es un lenguaje de programación dinámico de propósito general extremadamente popular. Éste lenguaje soporta múltiples paradigmas de programación como lo son la programación orientada a objetos, la programación funcional y la programación estructurada. Su filosofía hace énfasis en la legibilidad del código mediante el uso de indentación y se considera una de las herramientas principales en área de las ciencias de datos.

**2.4.2 Biblioteca Lifetimes**

Lifetimes es una biblioteca de software que implementa los modelos estadísticos Buy-Till-You-Die, Customer Lifetime Value y el modelo bayesiano PyMC (en Beta) en el lenguaje de programación Python. Es de código abierto bajo la licencia Apache 2.0*.*

**2.4.3 Biblioteca Pandas**

Pandas es una biblioteca de software escrita para el lenguaje de programación Python, comúnmente usada para la manipulación y el análisis de datos. Ofrece estructuras de datos y operaciones para manipular datos de forma tabular y series de tiempo. Es software libre publicado bajo la licencia BSD de tres cláusulas.

**2.4.4 Biblioteca MatplotLib**

Matplotlib es una biblioteca de gráficos para el lenguaje de programación Python. Se utiliza para incrustar gráficos en aplicaciones y poder visualizar los datos o resultados de modelos. Es de código abierto y se distribuye bajo la licencia de estilo BSD.

**2.4.5 Jupyter Notebook**

Jupyter Notebook (anteriormente IPython Notebook) es un entorno computacional interactivo muy popular en la ciencia de datos, el cual se ejecuta en la web y se utiliza para crear documentos que simulan un cuaderno digital. Cada documento contiene una lista ordenada de celdas de entrada/salida que pueden contener código de programación, texto (usando el formato Markdown), lenguaje matemático, diagramas e imágenes.

**2.4.6 Flask**

Flask es un framework de Python que se utiliza para crear aplicaciones y sistemas web. Este framework no requiere otro tipo de herramientas y bibliotecas, tampoco tiene componentes preexistentes lo que lo hace muy ligero. Es de código abierto bajo la licencia BSD y fue lanzado inicialmente en abril del año 2010.

**2.4.7 Javascript**

Javascript es un lenguaje de programación dinámico de alto nivel, es interpretado y su sintaxis se basa en los lenguajes C++ y Java. Todos los navegadores web modernos interpretan el código Javascript por lo que se considera el lenguaje de la web, se utiliza principalmente para añadir dinamismo a las páginas web en el lado del cliente, pero actualmente se utiliza también en el lado del servidor para crear aplicaciones de todo tipo.

**2.4.7 Typescript**

Typescript es un lenguaje de programación tipado, compilado y de código abierto bajo la licencia Apache. Fue desarrollado en Octubre del año 2012 y actualmente es mantenido por Microsoft como un subconjunto del lenguaje de programación Javascript.

El uso y popularidad de Typescript ha crecido exponencialmente en los últimos años debido a las mejoras que ofrece sobre el lenguaje Javascript, en especial la validación de tipos para las variables lo cual ayuda en la detección temprana de errores aumentando la robustez del software, se considera esencial para proyectos grandes.

**2.4.8 Vue.js**

Vue.js es un framework de código abierto escrito en Typescript que se utiliza en la construcción de interfaces de usuario dinámicas y aplicaciones web. Fue creado por Evan You en febrero del año 2014 bajo la licencia MIT. Este framework es de los más populares actualmente para crear interfaces web y se enfoca en una arquitectura de representación declarativa y composición de componentes.

**2.4.9 Nuxt**

Nuxt es un metaframework que utiliza como base a Vue.js, es de código abierto bajo la licencia MIT y tiene como objetivo facilitar la creación de aplicaciones web “universales”, es decir, sitios web que tengan la funcionalidad de una aplicación web y los beneficios de los sitios web estáticos. Fue creado en octubre del 2016 por Sebastien Chopin, Alexandre Chopin y Pooya Parsa.

**2.4.10 Chart.js**

Es una biblioteca gratuita de código abierto para la visualización de datos, está escrita en Javascript y fue creada en el año 2013 bajo la licencia MIT. Esta biblioteca permite mostrar datos en gráficos de barra, linea, área, circular, burbujas, radar, polar y de dispersión, todo esto utilizando la tecnología de HTML5 Canvas. Es una de las bibliotecas mas populares y se considera la mejor por su simplicidad de uso.