Autor: Dimitrio Mandamadiotis Kalfagiannis CI 23.723.158
Tutor: Yaneth Moreno

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

AVANCE DE PROYECTO DE GRADO

Introducción

El comercio electrónico o también conocido como e-commerce, consiste en la compra y venta de productos o servicios a través de internet, este comenzó a principios de 1970 y se desarrolló a mediados de la década de los 90 como una manera bastante sencilla de adquirir productos o servicios sin tener que desplazarse a una ubicación física.

La cantidad de comercio llevada a cabo electrónicamente ha crecido de manera extraordinaria y esto ha estimulado la creación y utilización de innovaciones como el marketing en internet, cadenas de suministros complejas, sistemas automáticos de recolección de datos y sistemas para la gestión de la relación con el cliente, estos últimos han permitido la aplicación de métodos cuantitativos para el análisis y la optimización del rendimiento en los e-commerce.

Una de las áreas más importantes de estudio es el comportamiento de los clientes, en ella se evalúa y se estudia la relación directa que existe entre el servicio que se les proporciona a los usuarios y los ingresos de la tienda de comercio electrónico, para esto se utilizan variables como el valor del ciclo de vida de un cliente y la tasa de churn y lo que se busca es obtener un incremento en las ventas mediante mejoras en los servicios, la experiencia del cliente y la calidad de los productos.

Antecedentes

En el mundo de los negocios y el comercio actual seguimos observando un crecimiento rápido de los sistemas digitales y las tecnologías de información, algunos de estos sistemas se basan en la gestión de relaciones con los clientes, los cuales son comunes en negocios contractuales como los servicios de telecomunicaciones o servicios de software basado en suscripción. El comercio electrónico es un negocio no contractual y en Europa tiene un crecimiento anual proyectado del 6 % hasta el año 2025, también en el año 2020 los ingresos en e-commerce aumentaron en un 10 % (Jílková & Králová, 2021), estas cifras nos demuestran lo vital que es para los negocios mantener a los clientes.

Durante los últimos tres años se han realizado muchos estudios sobre el comportamiento de los clientes en los negocios. (Falla, 2021) en su trabajo de grado *Predicción de Abandono de Clientes en Telecomunicaciones Mediante el Aprendizaje Automático,* utiliza técnicas y algoritmos de inteligencia artificial para predecir el abandono de clientes en un negocio con modelo contractual. También identificó varias técnicas de minería de datos para la identificación de clientes que están a punto de abandonar.

(Sinha & Raizada, 2022) en su trabajo *Modelling Customer Churn Rate and Its Use for Customer Retention Planning* muestran un modelo para un negocio contractual basado en máquinas de aprendizae o *machine learning* para predecir la tasa de abandono, adicionalmente dan una serie de recomendaciones a los negocios para mantener una tasa de abandono baja según sus descubrimientos.

La complejidad de modelar la tasa de abandono en negocios no contractuales viene de que no se conoce el momento en el que un cliente abandonó. Una manera de aproximar esta variable es mediante el valor de la compra media de un cliente. (Abdolvand, Albadvi & Koosha, 2021) mencionan en su artículo *Customer Lifetime Value: Literature scoping map, and an agenda for future research* que la literatura se enfoca en explicaciones teóricas para calcular el valor de la compra media de un cliente y no expresan las diferencias que pueden existir a la hora de aplicar el conocimiento a la práctica.

Planteamiento del problema

En un negocio como una tienda de comercio electrónico lo más importa son los clientes, sin ellos no hay ingresos y sin ingresos la tienda no podría existir. En la actualidad, el costo de atraer y convertir personas en nuevos clientes suele ser muy alto, además es mucho más fácil vender productos a un individuo que ya ha comprado antes en la tienda que a un desconocido, es por esto que vale la pena determinar cuándo los clientes podrían desertar para poder tomar acción antes y tratar de prevenir el abandono. Por otro lado, es esencial determinar con precisión cuántos de los clientes aún siguen siéndolo, con esta información se podría analizar el estado actual del negocio y proyectar escenarios para el futuro.

Calcular el churn de los clientes es más fácil en negocios contractuales, como proveedores de telefonía o de banda ancha, ya que es fácil ver cuándo estos están a punto de abandonar a medida que sus contratos se acercan a la fecha de finalización. Sin embargo, predecir la deserción en mercados no contractuales como el comercio electrónico es mucho más difícil porque no se conoce el momento de salida de un cliente, y en su lugar debe predecirse.

En este sentido, se considera que desarrollar un sistema capaz de predecir el churn de los clientes en un e-commerce sería de vital importancia, ya que permitiría analizar la salud de un negocio y le daría una ventaja a los administradores y encargados de área de marketing.

Justificación

El comercio electrónico a nivel mundial ha tenido un crecimiento exponencial en los últimos años y debido a esto se ha generado una cantidad gigantesca de datos que se pueden utilizar para optimizar métricas y alcanzar objetivos de negocio.

En los negocios contractuales es fácil establecer una base para analizar el comportamiento de los usuarios, esto porque se conoce con exactitud la cantidad de usuarios que dejaron de ser clientes en un momento dado, ya que no renovaron sus contratos o suscripciones. Estos datos permiten a los negocios contractuales crear sistemas complejos de predicciones de rendimiento y la optimización de la experiencia del usuario.

Dicho esto, hay otro grupo de negocios donde no se sabe con exactitud cuando un usuario deja de ser cliente. Es por ello que es necesario definir un sistema para predecir la deserción de los clientes en los negocios no contractuales como las tiendas de comercio electrónico. Este sistema le da la posibilidad a estos negocios de realizar análisis complejos e identificar problemas en las relaciones con sus clientes y tomar acción para prevenir el abandono a tiempo.

Objetivos

El objetivo general de este trabajo de grado es desarrollar un sistema capaz de inferir la deserción de los clientes en una tienda de comercio electrónico.

Para ello se debe cumplir con los siguientes objetivos específicos:

1. Construir un data Warehouse con información sobre patrones de comportamiento de los clientes en una tienda de comercio electrónico.

2. Definir el método estadístico apropiado para inferir la deserción de los clientes.

3. Desarrollar el modelo para inferir la deserción de los clientes en una tienda de comercio electrónico.

4. Validar el modelo desarrollado con un conjunto de datos de prueba.

5. Evaluar el rendimiento del modelo.

Metodología a utilizar

En la actualidad, la agilidad al cambio es un factor de suma importancia en los proyectos que involucren software. Los requisitos y diseños tienden a cambiar rápidamente con el tiempo para adaptarse a las necesidades del proyecto. Por esta razón, existen las metodologías ágiles, estas permiten darle un mayor enfoque al proceso de desarrollo, enfatizar la comunicación cara a cara en lugar de la documentación y en especial facilitar la refactorización y la adaptación a los cambios. En este proyecto se plantea aplicar el método Kanban combinado con un backlog, esto facilita la organización y optimiza el flujo del trabajo para la investigación y el desarrollo del sistema.

Kanban se basa en una estructura de flujo de trabajo continuo de forma visual. Los elementos de trabajo que son representados por tarjetas se organizan en un tablero de kanban, donde pasan de una etapa del flujo de trabajo (columna) a la siguiente. Las etapas habituales del flujo de trabajo son:

1. **Por hacer:** que representa el trabajo que no se ha empezado.
2. **En curso:** trabajo en el que se está trabajando activamente.
3. **En revisión:** trabajo que está finalizado y en espera de revisión.
4. **Finalizado:** trabajo completamente terminado.

Una característica importante de kanban es que se establece una cantidad máxima de trabajo que puede existir en cada estado del flujo de trabajo. Limitar la

cantidad de trabajo en curso mejora el rendimiento y reducen la cantidad de trabajo "prácticamente listo", ya que obliga al equipo a centrarse en un conjunto de tareas más pequeño, en este proyecto no se podrá trabajar en más de tres tareas activas. Esta característica puede ser una limitante a la hora de programar las tareas por hacer, es por ello que se va a tratar esta columna como un backlog.

El backlog es una lista de trabajo ordenado por prioridades para el equipo de desarrollo que se obtiene de la hoja de ruta y los requisitos. Los elementos más importantes se muestran al principio del backlog para que el equipo sepa qué hay que entregar primero. En este proyecto el backlog será atendido por el tutor y el autor.

Alcance

En el presente proyecto se plantea la creación de un sistema que analice los datos relacionados al comportamiento de los usuarios en un sitio web de comercio electrónico, algunos de estos datos incluyen vistas de productos, agregar o quitar productos de carritos de compra, añadir o quitar productos de favoritos y comprar. Estos datos servirán de entrada a un modelo que tendrá como salida la probabilidad de que el usuario abandonara la tienda.

Así mismo se hará uso del lenguaje Python para la implementación del modelo ya que es el lenguaje mas utilizado en el área de análisis de datos y la mayoría de las herramientas están implementadas en este lenguaje. La interfaz será web por lo tanto se usará el lenguaje Javascript para el frontend y el backend del sistema.

Desarrollo

En una primera fase realicé una investigación exhaustiva de los modelos utilizados para resolver este tipo de problemas y decidí tomar como base los modelos estadísticos Buy-Till-You-Die (BTYD), estos son una familia de modelos que tienen como objetivo describir el comportamiento de compra de los clientes como una serie de distribuciones. Existen varios modelos BTYD, pero todos tienen las siguientes características:

- **Modelado Probabilístico.** Todos estos modelos generan distribuciones 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.

- **Entrada de 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.

- **Distribuciones RFM.** Todos los modelos buscan capturar tres comportamientos de los clientes: Recency (recencia), Frecuency (frecuencia) y Monetary (Valor monetario).

De todos los modelos de la familia BTYD, seleccioné el modelo Pareto / NBD, su nombre se refiere a la distribución de Pareto utilizada para inferir la taza de deserción de clientes y la distribución binomial negativa (NBD por sus siglas en inglés) utilizada en la predicción de compras futuras. Éste modelo es el estándar en el campo para este tipo de análisis y además sirve como base al modelo "Abe", el cual puede aceptar como entrada datos que no están relacionados a los pedidos de los clientes, por ejemplo las vistas en el sitio web de los productos y la actividad de los carritos.

Ya al haber definido el modelo, realicé la segunda fase en donde busqué una base de datos pública para entrenar y probar el modelo, esta base de datos debía contener una tabla de pedidos y un histórico de al menos un año. Mi búsqueda resulto en dos bases de datos públicas dentro de la plataforma **Kaggle** que cumplen con los requisitos de los modelos BTYD y que puedo usar en el proyecto:

- Conjunto de datos públicos de comercio electrónico brasileño por Olist: ésta base de datos contiene alrededor de 100.000 pedidos realizados en la tienda Olist durante los años 2016 y 2018. Es data comercial real anónima. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

- Minorista del Reino Unido. Se trata de un conjunto de datos que contiene todas las transacciones que se produjeron durante los años 2010 y 2011 para una tienda minorista e-commerce con sede en el Reino Unido. La empresa vende principalmente regalos únicos para toda ocasión. Muchos clientes de la empresa son mayoristas. https://www.kaggle.com/datasets/carrie1/ecommerce-data

Como tercera fase definí la arquitectura del sistema, para ello decidí encapsular el modelo que genera la información dentro de un backend para que éste pueda ser consumido por una aplicación web. El backend se realizará en Python ya que es el mismo lenguaje donde se implementará el modelo. El frontend se llevará a cabo con el lenguaje Javascript ya que es el lenguaje de la web y con él está escrito uno de los frameworks que manejo llamado Vue.

Actualmente estoy trabajando en la implementación del modelo en Python y el diseño de la interfaz de usuario. También estoy comenzando el capitulo 2 del documento del proyecto de grado.

Bibliografía

Sinha, A. & Raizada, S. (2022). Modelling Customer Churn Rate and Its Use for Customer Retention Planning. http://dx.doi.org/10.2139/ssrn.3998408

Jílková, P. & Králová, P. (2021). Digital Consumer Behaviour and eCommerce Trends during the COVID-19 Crisis. https://doi.org/10.1007/s11294-021-09817-4

Rehkopf, M. (2022). What is a kanban board? https://www.atlassian.com/agile/kanban/boards

S. Wu, W. -C. Yau, T. -S. Ong & S. -C. Chong. (2021) Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. https://ieeexplore.ieee.org/abstract/document/9406002

Falla, J. (2021). Predicción de Abandono de Clientes en Telecomunicaciones Mediante el Aprendizaje Automático. http://hdl.handle.net/20.500.12010/22247

Wackerly, D., Mendenhall, W. & Scheaffer, R.L. (2014). Mathematical Statistics with Applications.

Gold, C. (2020) Fighting Churn with Data.

Abdolvand, N., Albadvi, A. & Koosha, H. (2021) Customer Lifetime Value: Literature scoping map, and an agenda for future research.

Schmittlein, D., Morrison, D., & Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next?

Fader, P., Hardie, B., & Lee, K. (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model.

Abe, Makoto. (2009) Counting your customers one by one: A hierarchical Bayes extension to the Pareto/NBD model.

Palabras Claves

1) Comercio Electrónico
2) Predicción Churn
3) Inteligencia Artificial
4) Machine Learning
5)