SEGMENTACIÓN DE USUARIOS PARA MEJORA DE EXPERIENCIA EN SITIOS WEB DE COMERCIO ELECTRÓNICO

¿Cómo podemos identificar los factores principales que llevan a los usuarios a abandonar el proceso de compra para optimizar estrategias de mercadeo y retención?

Grupo:

Tomas Vargas, Valentina Colonna, Erwin Cañon, Jeremy Quintero

Universidad de los Andes

Maestría en Inteligencia Analítica de Datos

Proyecto Aprendizaje No Supervisado

2023

Resumen

El comercio electrónico se ha vuelto esencial para las empresas en la era de internet y la globalización, pero comprender y mejorar la retención de clientes en las páginas web es un desafío complejo. Este proyecto busca construir un modelo de clustering no supervisado para segmentar a los usuarios en grupos basados en su comportamiento en el sitio web. El objetivo no es predecir ventas, sino identificar los factores que llevan a los usuarios a abandonar el proceso de compra. Se utilizarán modelos como DBSCAN, Clustering Jerárquico y K-medias para agrupar a los usuarios y analizar dónde se detienen en el embudo de conversión. Estos grupos servirán como insumo para optimizar estrategias de marketing y retención de clientes, permitiendo a los especialistas del comercio electrónico identificar características clave. Aunque se considera DBSCAN debido a sus ventajas, se realizarán comparaciones con otros métodos de clustering para tomar una decisión informada. El proyecto busca responder a la pregunta de cómo identificar los factores principales que afectan el abandono de usuarios en el proceso de compra y, así, mejorar las estrategias de mercadeo y retención.

Introducción

Con el alza del internet y la globalización, el comercio electrónico se ha convertido en una parte vital de los negocios para alcanzar a sus usuarios, dar a conocer sus propuestas de valor y ofrecer sus productos y/o servicios para una posible venta. La construcción de estas herramientas y el seguimiento de estas es un problema complejo debido al desarrollo que conllevan y los diferentes factores que pueden determinar que una compra se lleve a cabo o no. Para los comercios, es de gran interés conocer cuáles son estos factores que llevan al usuario a realizar o no la compra de algunos de los productos y/o servicios ofertados. Por ejemplo, si el negocio desea aumentar la retención de los clientes en las páginas web que tiene para aumentar sus ventas, ¿cómo podemos identificar los factores principales que llevan a los usuarios a abandonar el proceso de compra? Teniendo en cuenta que muy probablemente se tienen recursos limitados para invertir en esta mejora en la retención, ¿cómo se pueden optimizar estos recursos enfocando su uso en los factores que tienen mayor impacto en el abandono de los usuarios? Con esto en mente, el propósito de este proyecto es construir un modelo que agrupe estos usuarios con base en su comportamiento en el sitio web, lo cual puede ser posteriormente utilizado para análisis de funnel de conversión identificando donde se detienen los usuarios en el embudo de conversión, lo cual es vital para optimizar estrategias de marketing y aumento de retención de clientes.

Teniendo en cuenta la problemática planteada y la forma como se plantea solucionar, para este proyecto vamos a hacer uso de modelos de clustering no supervisados para segmentar a los usuarios que usaron la página web. Recordemos que a pesar de que el objetivo final de un ecommerce es lograr una venta, en nuestro caso no estamos interesados en desarrollar un modelo de clasificación que nos permita identificar cuáles son los usuarios que van a terminar en una venta o no con base en sus características. En nuestro caso, nuestro objetivo es segmentar a los clientes en múltiples grupos por características similares con base en su comportamiento de navegación y así, que estas agrupaciones puedan ser usadas por el ecommerce como insumo y guía en la optimización de estrategias de marketing y retención de clientes.

Este es un problema ampliamente abordado por su interés y utilidad para los ecommerce, ya que es imperativo conocer a tus usuarios y agruparlos por características comunes para el desarrollo de estrategias. Este es un problema ampliamente abordado por su interés y utilidad para los ecommerce, ya que es imperativo conocer a tus usuarios y agruparlos por características comunes para el desarrollo de estrategias.

Por ejemplo, en la investigación “Machine-Learning Techniques for Customer Retention: A Comparative Study” [https://pdfs.semanticscholar.org/2a9f/505e1ab148aa3d91810f509ee133272be554.pdf] llevan a cabo pruebas con diversos modelos de clasificación supervisados con el fin de presentar cuales muestran un mejor desempeño prediciendo si un usuario va a abandonar la compra o no. De forma similar, el trabajo que se va a llevar a cabo en este proyecto es la implementación de un modelo para la mejora en la retención de clientes en E-Commerce, la diferencia radica en el enfoque, ya que en nuestro caso desarrollaremos un modelo de clustering no supervisado que permita agrupar a los usuarios por sus características principales y que esto sirva de insumo para expertos en futuras diseños e implementaciones de estrategias de marketing.

Materiales y Métodos

Para abordar esta pregunta, utilizamos el conjunto de datos "Online Shoppers Purchasing Intention Dataset" disponible en el siguiente link . Estos datos constan de 12.330 registros y 18 variables, que incluyen información sobre el comportamiento del usuario en el sitio web, su duración en las páginas, el tipo de tráfico, la región y otras variables a considerar.

Primero, realizamos un proceso de limpieza de datos en donde verificamos los valores nulos. Luego, transformamos las variables de Revenue y Weekend que originalmente tenían valores 'true' y 'false' en valores numéricos, asignando 1 y 0. Continuando, convertimos las variables de categóricas a variables dummies y obtuvimos las estadísticas descriptivas de cada una de las variables. En las estadísticas descriptivas podemos observar que por sesión los usuarios entran en promedio a 2 páginas administrativas, la mitad de ellos visitó por lo menos 18 páginas relacionadas a productos y las tasas máximas de rebote y salida que tenemos en alguna(s) pagina(s) es 20%. Mientras que en los histogramas podemos ver de forma rápida cómo se distribuyen los valores para cada variable. Por ejemplo, vemos que, aunque la mayoría de las ExitRates se acumulan entre 0 y 0.05 (Gráfico 1), tenemos un pico importante en 0.2, así mismo vemos que la mayoría de los datos el PageValue es 0, esto tiene sentido teniendo en cuenta que la mayoría de las páginas no es donde se realiza la compra como tal, ya que muchas pueden ser informativas. Luego, realizamos un gráfico de correlación (Gráfico 2) donde revisando las correlaciones, como era de esperarse, aquellas que saltan a la vista es la duración en la paginas con respecto al número de páginas que se visitan de esa misma categoría, así como la relación entre sí hay una compra o no y el valor de la página, ya que recordemos que el valor de la página se calcula teniendo en cuenta si en esa sesión se realizó alguna compra.

Gráfico 1. Histogramas

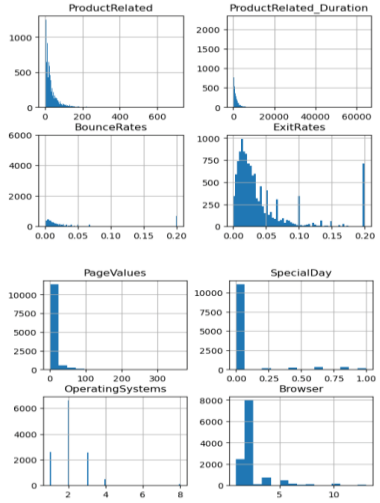
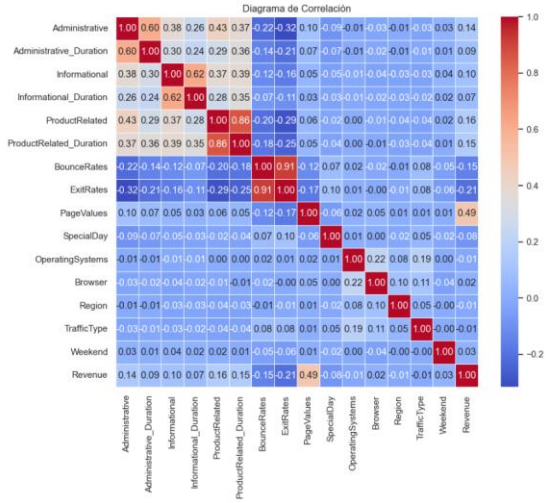


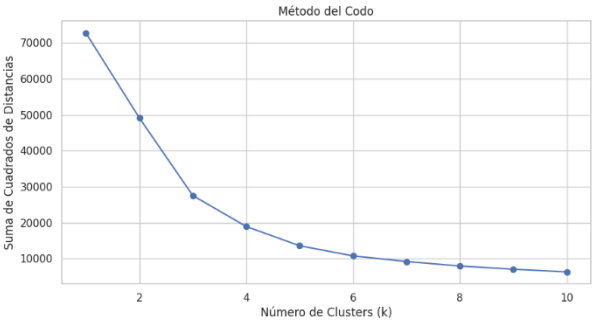
Gráfico 2. Matriz de correlación

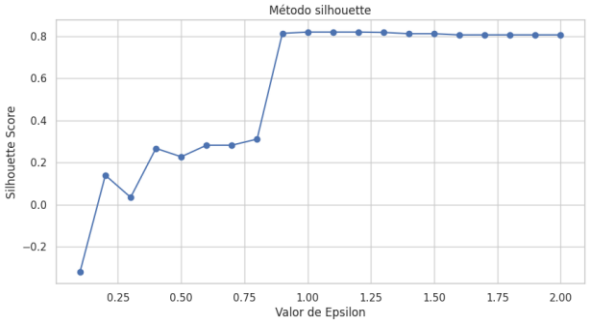


Luego, estandarizamos los datos y aplicamos diferentes modelos de aprendizaje no supervisado, para segmentar los usuarios de acuerdo con su comportamiento en el sitio web. Los algoritmos que utilizamos fueron Clustering jerárquico, K-Means, DBSCAN y por último Isolation Forest. La implementación y los resultados se presentan en la sección siguiente.

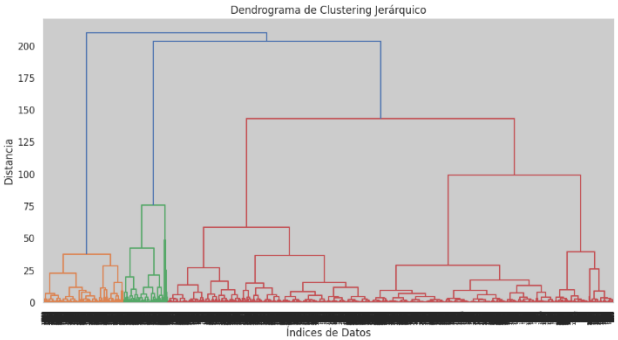
Resultados y Discusión.

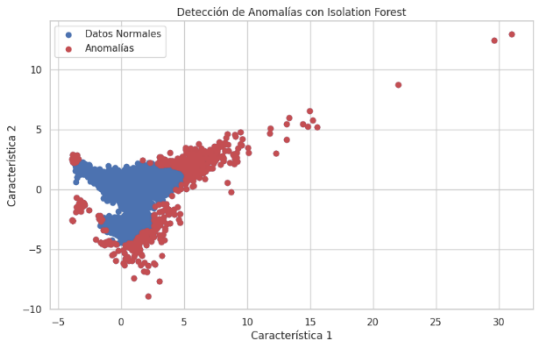
A continuación, describiremos los pasos que se llevaron a cabo durante la implementación de los algoritmos y los parámetros seleccionados en cada uno. Para empezar, se redujo la dimensionalidad utilizando PCA con el objetivo de tener el gráfico segmentado en dos componentes. A continuación, utilizamos el método elbow para determinar la cantidad de clusters (3) y el método silhouettey el valor de epsilon (0.99) preliminar que se iba a utilizar.





Luego, se construyó un modelo de clustering jerárquico utilizando el método de enlace 'ward' y se configuró inicialmente con 3 clusters, por último, se utilizó un dendrograma para visualizar los resultados. Continuando, se construyó un algoritmo K-means configurado para 3 clusters y una semilla aleatoria. Para el tercer algoritmo, se construyó un modelo de K-medoides en donde se calculó la matriz de distancias entre los puntos de datos estandarizados usando la distancia euclidiana y luego configuró K-Medoids para agrupar los datos en 3 clusters. Para el cuarto algoritmo DBSCAN, se determinó el eps y los min sample óptimos resultantes en 0.106 y 25 respectivamente. Luego se aplicó el algoritmo con esos parámetros y se graficó la agrupación DBSCAN. Por último, se construyó el modelo de Isolation Forest el cual nos muestra la segmentación de los datos en anomalías y datos normales. A continuación, se muestran varias de las gráficas obtenidas para Clustering Jerárquico e Isolation Forest:





Los resultados de nuestro análisis nos muestran la existencia de 3 segmentos distintos de clientes en línea. Estos segmentos se caracterizan por diferentes patrones de compra y comportamientos. Podemos ver que para los clientes del cluster 1, el promedio de duración en páginas administrativas es de 19 segundos y su promedio de salida es de 0.06 y el bounce rate de 0.03. Mientras que para el segundo cluster, el promedio de duración en páginas administrativas es de 87 segundos y su promedio de salida

es de 0.01 y el bounce rate de 0.02. Por último, en el tercer cluster el promedio de duración en páginas administrativas es de 196 segundos y su promedio de salida es de 0.02 y el bounce rate de 0.006.

	Cluster	0	1	2
ExitRates	count	6960.000000	1768.000000	3602.000000
	mean	0.061133	0.018698	0.020140
	std	0.057043	0.020205	0.012200
	min	0.000000	0.000000	0.000000
	25%	0.022222	0.006897	0.011753
	50%	0.040000	0.013333	0.018076
	75%	0.075000	0.024419	0.026029
	max	0.200000	0.200000	0.106667
BounceRates	count	6960.000000	1768.000000	3602.000000
	mean	0.035424	0.002541	0.006267
	std	0.060679	0.013441	0.007989
	min	0.000000	0.000000	0.000000
	25%	0.000000	0.000000	0.000000
	50%	0.008333	0.000000	0.003867
	75%	0.034737	0.000000	0.008962
	max	0.200000	0.200000	0.080838
Administrative_Duration	count	6960.000000	1768.000000	3602.000000
	mean	19.104505	87.079706	196.993112
	std	49.004624	152.725908	264.628967
	min	0.000000	0.000000	0.000000
	25%	0.000000	0.000000	47.017857
	50%	0.000000	46.400000	119.366667
	75%	11.000000	110.500000	243.125000
	max	844.000000	1946.000000	3398.750000

Conclusión

- De acuerdo con los resultados obtenidos en los diferentes modelos, podemos determinar que algunos de ellos presentan información que no es fácil de interpretar, como por ejemplo Clustering Jerárquico.

- Vemos que el modelo de K-medoides logra construir diversos grupos de usuarios similares, lo cual permitiría a los comercios dividir a sus usuarios por características similares, lo cual es un gran insumo para el diseño de estrategias significativas en mejoras de retención de usuarios en los procesos de compra que es el objetivo del uso de los modelos de aprendizaje no supervisado.

- Por último, esta estrategia y este modelo son replicables en diversos comercios E-Commerce con recolección de datos similares. Esta flexibilidad permite al modelo ser fácilmente implementado en diversas industrias con este mismo problema de retención de clientes.

- Adicional podemos observar que los tiempos de duración en las páginas son una variable relevante asociada a la retención de clientes.

Se recomienda para el cluster 1, que son los clietnes que más tiempo están en la página, se realice un estudio en detalle para generar estrategias de marketing que permitan materializar la retención.
Para el cluster 2 se recomienda realizar campañas con boletines informativos, registrarse para eventos web o participar en encuestas.
Para el cluster 3 sería conveniente ofrecer un programa VIP que permita su permanencia.

Bibliografía

Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016, October 1). *Review on Customer Segmentation Technique on Ecommerce*. Latest TOC RSS. https://www.ingentaconnect.com/contentone/asp/asl/2016/00000022/00000010/art00090

Sabbeh, S. F. (2018). *Machine-Learning Techniques for Customer Retention: A Comparative Study* (thesis). https://pdfs.semanticscholar.org/2a9f/505e1ab148aa3d91810f509ee133272be554.pdf

Sakar, C., & Kastro, Y. (2018, August 30). *Online shoppers purchasing intention dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Google. (n.d.). *Exit rate vs. Bounce Rate - analytics help*. Google. https://support.google.com/analytics/answer/2525491?hl=en