Definición de productos en eventos promocionales en Makro para maximizar la rentabilidad en las promociones.

Mariana Árias, Juan Camilo Valencia, Esteban Moreno Cediel, Viviana Vales.

Resumen

Se busca definir los productos que se ofertarán en las diferentes promociones con el fin de lograr la mayor fidelidad del cliente (mayor atracción de la promoción para cada tipo de cliente) sin sacrificar el lucro de la venta. Para ello, utilizamos una base de datos que contiene las ventas, el lucro, las cantidades vendidas y el número de facturas acumuladas del 2022 para cada tipo de cliente y cada producto. Finalmente, se decidió implementar un PCA para reducir la dimensionalidad del problema y se utilizó un sistema de recomendación basado en contenidos para recomendar los productos que deberían ser promocionados dependiendo de cada tipo de cliente. Como resultado principal, se obtiene las recomendaciones basados en similitud utilizando CountVectorizer, donde se obtienen los 10 productos que se deberían de promocionar con mayor agresividad por cada tipo de cliente, con el fin de aumentar y fidelizar el mayor numero de clientes posibles.

Introducción

¿Cuáles son los artículos que maximizan la ganancia de la compañía perteneciente al sector retail en un evento promocional? Makro es una cadena mayorista multinacional holandesa del sector privado que tiene más de 100 empleados. Esta empresa realiza 2 eventos promocionales (Economakro y Reventon) 2 veces al mes cada uno los cuales están enfocados a tipos de clientes específicos (clientes profesionales para Economakro y clientes individuales para Reveton). El reto de cada evento promocional es poder dar la mejor oferta en una cantidad determinada de artículos, que logre atraer el mayor número de clientes a las tiendas, maximizado la ganancia de la empresa. Uno de los principales obstáculos es poder determinar los artículos que se deben ofertar en cada evento promocional y el descuento máximo permitido por artículo para no sacrificar el lucro de la compañía. Esta decisión es importante ya que, si la única variable de decisión para escoger los productos que se deben ofertar fuera la venta, se escogería el azúcar, la leche en polvo y la cerveza. Sin embargo, estos productos no tienen un buen margen de ganancia por lo que ofrecerlos en promoción resulta inviable. Además, hay que entender el comportamiento del consumidor ya que un cliente corporativo (estaciones de gasolina) compra productos y cantidades muy diferentes a un cliente individual. Se encontró literatura nacional e internacional que trata sobre problemas y contextos muy similares al nuestro. Sin embargo, a pesar de la importancia que tiene descifrar esta combinación perfecta dentro de la investigación q realizamos y los proyectos que encontramos el foco principal era encontrar combinaciones de productos a nivel general sin distinguir entre productos por tipo de cliente y utilidad. Una vez realizado el ejercicio, se obtiene que para los modelos basados en cluster, sea k-medias o DBSCAN, los resultados no son los esperados, y el modelo se vuelve ineficiente debido al gran numero de variables categóricas que se deben escalar, dando un numero de cluster muy alto, con resultados poco eficientes y poco interpretables. Por otro lado, al utilizar el método de recomendaciones basado en similitud, se obtienen los 10 articulos mas recomendados para generar promociones según el evento promocional, sin embargo, este método, por la estructura de la base, tiene un gran gasto computacional, generando errores en computadores con poca capacidad en cuanto a hadware y software, o generando grandes demoras para obtener un resultado deseado, por lo que se recomienda buscar un método mas acorde con el problema.

Materiales y métodos

Nuestra base de datos se compone de los registros de ventas acumuladas del año por producto. Según las estadísticas descriptivas contamos con 168,257 registros sin ningún valor nulo. De las columnas que tenemos en nuestra base de datos tenemos 3 de tipo Object (Grupo de clientes NAME, Departamento

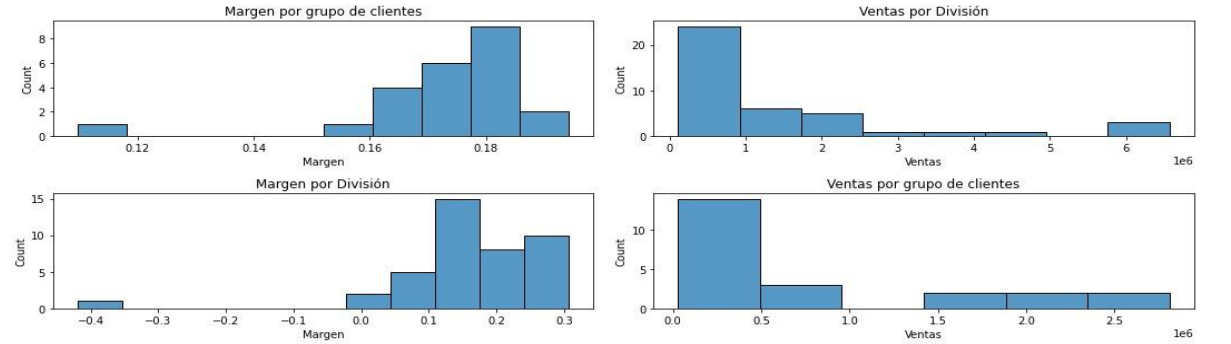
Name, Articulo Name) y 4 de tipo float (Articulo NR, Ventas, Ctd vendida, Lucro, Clientes activos, Ctd facturas y margen). Encontramos que esta base de datos contiene información sobre 16,637 productos diferentes. Luego, construimos un correlograma de nuestra base arrojando que las siguientes variables *Clientes Activos – Cantidad de Facturas*, y *Lucro – Ventas* tienen una alta correlación entre ellas. Teniendo en cuenta lo anterior, se deciden eliminar los registros que contengan al menos una de las siguientes condiciones:

- Corresponde al departamento llamado Service o Liquid porque son costos asociados al servicio de distribución, gestión de créditos, gasolina o biodiesel.
- Tiene cantidades vendidas menores o iguales a 0 porque puede corresponder a costos de distribución que son costos de transporte que asume la empresa
- Tiene ventas menores o iguales a 0 ya que pueden ser devoluciones
- Se eliminará la variable o columna Clientes activos ya que contiene exactamente la misma información que la variable Tipo de cliente. Por lo tanto, solo usaremos la variable tipo de cliente para nuestros análisis

Después de limpiar la base, quedamos con 168,078 registros que contienen la información de 16,588 productos diferentes. Por lo tanto, los diagramas de distribución y estadísticas descriptivas de cada variable de la base de datos limpia se presenta a continuación:

```
df.describe()
```

	Artículo NR	Ventas	Ctd Vendida	Lucro	Clientes Activos	Ctd Facturas	Margen
count	168078.000000	1.680790e+05	168078.000000	1.680780e+05	168078.000000	168078.000000	168078.000000
mean	190515.775580	1.130542e+06	103.511204	1.277825e+05	24.229548	40.939433	0.173938
std	190854.182324	1.777108e+07	1994.102534	1.221823e+06	131.770717	201.635412	0.176903
min	486.000000	9.000000e+00	0.001000	-4.259869e+07	1.000000	1.000000	-11.318848
25%	116329.000000	1.949000e+04	2.000000	2.509065e+03	1.000000	1.000000	0.120339
50%	129047.000000	6.444100e+04	6.000000	9.823956e+03	4.000000	4.000000	0.194447
75%	138016.000000	3.101800e+05	26.000000	4.933044e+04	12.000000	15.000000	0.262311
max	999674.000000	2.776281e+09	498274.000000	2.044718e+08	10797.000000	23024.000000	1.000000



Implementación del algoritmo.

Una vez estudiada a fondo nuestra base de datos damos paso a la implementación del algoritmo que nos ayudara a determinar cómo se agrupan los productos que adquieren los diferentes clientes de la organización.

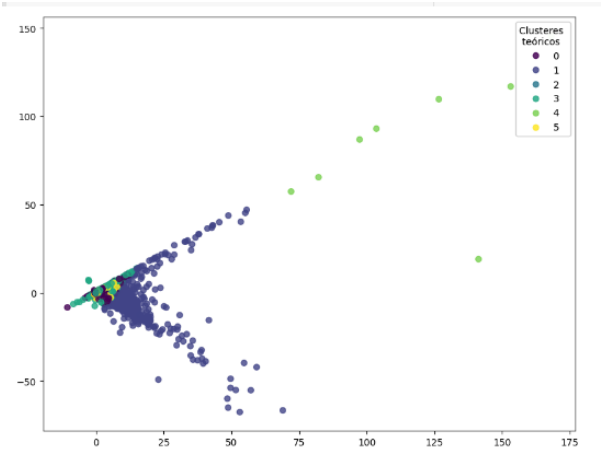
Para ello el primer paso es transformar cada una de las variables dicótomas con las que contamos, posteriormente escalamos la base con el fin eliminar el ruido que podrían tener las diferentes medidas y así obtener un mejor resultado en nuestro modelo de ML.

Con nuestra base escalada procedemos a aplicar PCA que nos permitirá reducir nuestra matriz de información capturando un gran porcentaje (>75%) de la varianza de nuestra base de datos capturando 2 componentes principales, de esta forma ya tenemos nuestra base lista para proceder a aplicar los distintos clústeres.

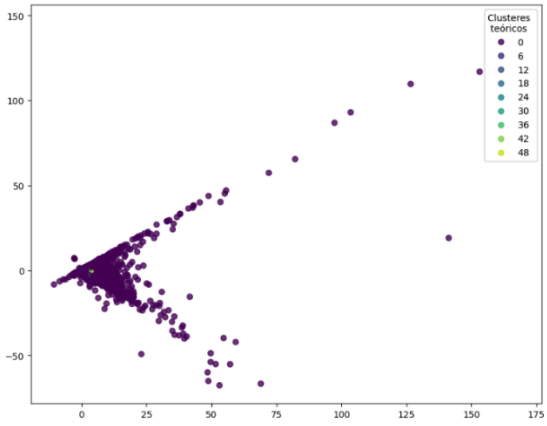
En este caso aplicaremos K-medias que es un algoritmo de clasificación no supervisada (cauterización) que agrupa objetos en k grupos basándose en sus características, este se caracteriza por su bajo costo computacional y por basar su funcionamiento en ubicar los centroides de los clúster en las distancias promedio de las observaciones lo que lo vuelve un poco ineficiente ante la presencia de datos atípicos. También aplicaremos DBSCAN que es un algoritmo basado en la densidad de las observaciones y que a diferencia de K-medias nos permite identificar los datos atípicos.

Revisamos la aplicación de los algoritmos planteados y refinamos el proceso de aplicación de k-medias y DBSCAN, obteniendo los siguientes resultados.

Para k.means se definió trabajar con 6 clústeres, obteniendo el resultado que aparece en la gráfica 1.1 ubicada a la izquierda. Para DBSCAN observamos como el algoritmo arrojo la creación de una gran cantidad de clúster teóricos (Gráfica 1.2), pero en los puntos solo podemos apreciar un clúster bien definido



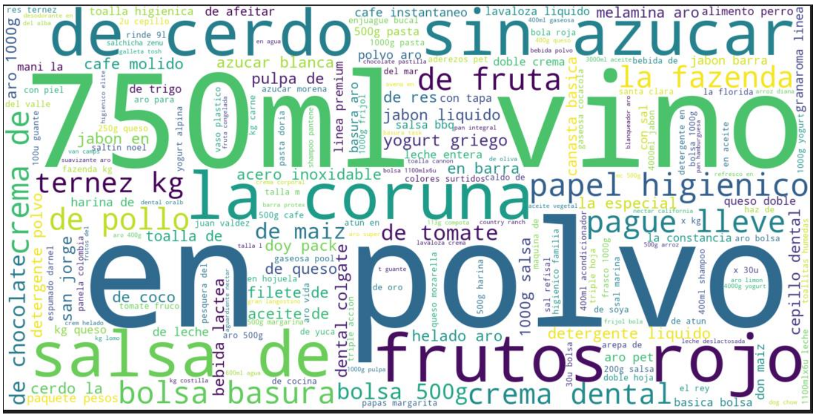
Gráfica 1.1.



Gráfica 1.2

Tras aplicar el algoritmo de k-medias encontramos una agrupación de los datos que permite observar algunos clústeres definidos, mientras que al aplicar DBSCAN observamos como el algoritmo arrojo la creación de una gran cantidad de clúster teóricos, pero en los puntos solo podemos apreciar un clúster bien definido.

Tales resultados, poco interpretables y poco consistentes, nos llevan a considerar un error en la metodología planteada en la creación del clúster, y que estos métodos no son los mas apropiados debido a la gran cantidad de variables categóricas que tiene la base, por lo cual se procede a realizar un algoritmo basado en sistema de recomendaciones, teniendo como tipo de motor la popularidad del artículo en cuanto al tipo de cliente que lo lleva y la frecuencia de compra que tuvo cada uno de ellos, con el fin de identificar que productos se deberían promocionar, según el tipo de cliente al cual esté enfocado el evento.
Como primer paso, se realiza un perfilamiento de la base de datos, eliminando los stopwords, eliminando números innecesarios en los artículos, caracteres especiales, palabras repetidas e innecesarias (Impoconsumo, Item, Delivery), urls si es que las hay, lematizando y tokenizando los artículos, se tiene entonces la siguiente nube de palabras:



Se procede entonces a tokenizar, lematizar y eliminar stopwords, para mejorar el análisis siguiente.

Una vez realizado este procedimiento, se realiza el sistema recomendador, Basados en similitud utilizando CountVectorizer, el cual utiliza la distancia de coseno, con respecto a los artículos mas frecuentados por los clientes, según el tipo clasificatorio de estos mismos. Tomando como ejemplo tres tipo de clientes, el resultado fue el siguiente:

Bares-Discotecas	Aceite De Soya Country Ranch 20l
	Whisky Buchanans Master 750ml
	Aceite Vegetal Country Ranch 20l
	Aguardiente Antioqueno Sin Azucar Botella 750ml
	Cerveza Coronita Extra 210mlx6u
	Cerveza En Lata Aguila Original 330mlx6u
	Aguardiente Nectar Club Botella 375ml
	Cerveza Corona Botella 355ccx6u
	Whisky Buchanans 18 Anos 750ml
	Aguardiente Antioqueno Verde Sin Azucar 24 750ml

Centros Med. Hospitales	Papel Reprograf Fotocopia Carta 75g
	Aceite Vegetal Country Ranch 20l
	Arroz Don Perfecto 50kg
	Aceite De Soya Country Ranch 20l
	Papel Reprograf Fotocopia Oficio 75g
	Arroz Caribe Vita 50kg
	Arroz Roa Fortiplus 500gx25u
	Leche Deslactosada Alqueria Bolsa 1100mlx6u
	Cafe Molido Sellorojo Fuerte 2500g.
	Huevo Amarrado Ta M&C Tipo A 30u

Consumidor Individual	Gaseosa Coca Cola 1.5l
	Gaseosa Coca-Cola Sabor Original 3lx6u
	Cerveza Coronita Extra 210mlx6u
	Huevo Amarrado Ta Aro Tipo Aa 30u
	Aceite Vegetal Country Ranch 20l
	Arroz Roa Fortiplus 500gx25u
	Cerveza En Lata Aguila Original 330mlx6u
	Paca Arroz Florhuila 500gx25u
	Cerveza Corona Botella 355ccx6u
	Papel Higienico Familia Acolchamax Megarollo 31m 24u

En las tablas anteriores, se puede observar el resultado de las recomendaciones de los artículos que se deberían de dar con promociones mas agresivas, según el tipo de cliente que clasifica el establecimiento comercial. Recordemos que son 19 tipos de clientes diferentes, por lo que se podría realizar recomendaciones para cada uno de ellos.

Conclusiones.

Los resultados obtenidos muestran que para utilizar métodos basados en cluster (K-Means y DBSCAN) se debe tener un tratamiento especial para las variables categóricas, ya que los resultados pueden ser poco contundentes, o sin una interpretación valida. En este caso, la gran cantidad de variables categóricas que presenta la base, es decir, cada articulo que existe en el portafolio de la compañía, hace que los resultados al aplicar este tipo de algoritmos basados en cluster sean poco eficientes, dando una gran cantidad de clusters teóricos, que para el objetivo final no aportan en nada. Por otro lado, utilizar el análisis de componentes principales ayuda en gran medida a simplificar el análisis, debido al alto numero de variables cuantitativas que tenia la base, al reducir su numero de 6 variables a 2, conlleva a que el análisis se realice de una manera mas eficiente, y el gasto computacional sea menor.

El otro algoritmo utilizado para este informe fue un sistema de recomendaciones basados en similitud utilizando CountVectorizer, en el cual, luego de dejar la base de artículos tokenizada y lematizada, obtenemos los 10 artículos más recomendados por cada tipo de cliente, y por las variables resultantes del PCA inicial, esto permite a la compañía realizar campañas promocionales mas agresivas según el tipo de cliente al que se realiza cierto evento, permitiendo la fidelización de clientes antiguos, como la atracción de clientes nuevos, aumentando la ganancia de la compañía misma. Este algoritmo sin embargo, dada la estructura y tamaño de la base, requiere un gran gasto computacional, y un gran tiempo de procesamiento, lo que hace que este algoritmo tampoco seal el mas eficiente para enfrentar este tipo de problemas, por lo cual, se recomienda en futuros análisis, realizar un tratamiento diferente a la base, o realizar un algoritmo diferente que permita obtener resultados mucho más concisos, y eficientes computacionalmente.

Bibliografía:

Bellini, P., Palesi, L.A.I., Nesi, P. *et al.* Multi Clustering Recommendation System for Fashion Retail. *Multimed Tools Appl* (2022). https://doi-org.ezproxy.uniandes.edu.co/10.1007/s11042-021-11837-5

Richard Weber, C.B (2017). Clustering retail producs based on customer behavior. En SI:Applied Soft Computing For Business Analytics (pags. 752-762). https://www-sciencedirect-com.ezproxy.uniandes.edu.co/science/article/pii/S1568494617300728?via%3Dihub#bib0045

I-Fei, C., Chi-Jie, L. (2021). Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering and Machine Learning. Japan. https://www-proquest-com.ezproxy.uniandes.edu.co/docview/2576497290?pq-origsite=primo&accountid=34489