Entrega propuesta inicial del proyecto

RESUMEN

El presente trabajo tiene como objetivo identificar cuáles son los productos que maximizan la ganancia de una compañía perteneciente al sector retail en un evento promocional haciendo uso de técnicas de aprendizaje no supervisado. Bajo este objetivo se implementarán técnicas de reducción de información (PCA) y técnicas de segmentación de clientes (K-means), que nos permitirán encontrar que tipo de clientes demandan en mayor medida determinados productos y como la combinación de estos logran maximizar las ganancias de la compañía.

Palabras clave.

Aprendizaje no Supervisado, PCA, K-Means, Sector Retail.

INTRODUCCIÓN

Uno de los puntos más importantes a la hora de entender el comportamiento de los clientes con el objetivo de mejorar las utilidades de la compañía es la información sobre sus hábitos de consumo, en este sentido, encontrar la mejor combinación de productos que podría contener una oferta que maximizara la utilidad del cliente y a su vez maximizara las ganancias de la compañía es una uno de los grandes retos que enfrenta cualquier empresa del sector retail.

Sin embargo, a pesar de la importancia que tiene descifrar esta combinación perfecta en muchas compañías aún no existe una técnica adecuada para la selección de los productos para eventos promocionales que distinga entre productos por tipo de cliente y utilidad.

El problema suele volverse más complejo cuando las compañías a pesar de acumular una gran cantidad de información sobre los clientes no resultan capaces de hacer uso de esta y suelen terminar atestados de data que confunde los tomadores de decisiones y consumo recursos de almacenamiento.

Partiendo de lo anteriormente expuesto este trabajo tiene como objetivo aplicar técnicas de aprendizaje no supervisado que permitan mediante técnicas de reducción de información y segmentación de grupos desarrollar un modelo que permita a cualquier compañía (en este caso Makro Colombia) identificar qué tipo de productos permiten maximizar la utilidad de un determinado tipo de cliente al encontrarlos en un paquete promocional a la vez que logran maximizar la utilidad de la compañía.

REVISIÓN PRELIMINAR DE LA LITERATURA.

Articulo/Proyecto/Investigación	Similitudes con el enfoque propuesto en esta entrega	Diferencias del enfoque propuesto en esta entrega
Clustering retail products based on customer behavior	-Se enfoca en el sector retail -Busca generar grupos/clusters de productos de diferentes con base en el comportamiento de los compradores -El análisis es utilizado para definir la mejor combinación de productos en las promociones -El parámetro de numero de clusters es definido a priori -La información es recolectada con base en los recibos de venta que tiene el establecimiento	-Sugiere un nuevo método para encontrar los clusters de productos: Un modelo de optimización en donde penaliza que un cliente compre 2 o más productos de una misma categoría y la función objetivo es minimizar las decisiones erróneas (que un cliente compre 2 productos de una misma categoría en una promoción)
Multi clustering Recommendation System for Fashion Retail	-Utiliza el algoritmo de k medoiodes para encontrar diferentes agrupaciones/clustering de productos en el sector retail de la moda -Utiliza el método de Silhouette para calcular el numero óptimo de clusters a encontrar. -Encuentra los clusters basándose en el comportamiento del consumidor. -El estudio fue hecho en Florencia, España. -La información es recolectada con base en los recibos de venta que tiene el establecimiento	-Se enfoca en el sector de la moda -Encuentra clusters de productos, pero de clientes también.
Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering	- Utiliza la distancia euclidiana para medir la distancia entre clusters. - Busca generar grupos/clusters de productos de diferentes con base en el comportamiento de los compradores -Divide la data en sets de entrenamiento y prueba.	-Se enfoca en el sector retail de la moda - El objetivo del clustering es para predecir la demanda únicamente y no con objetivos comerciales. -Encuentra dichos clusters dividiendo los artículos en venta rápida, media o demorada.

DESCRIPCIÓN DE LOS DATOS

Estadísticas descriptivas

La fuente de información proviene del banco de datos de la empresa, con el debido permiso y confidencialidad de estos. Contamos con una base con una temporalidad de 7 meses del año 2022, con 168.568 registros la cual tiene las siguientes variables:

- **Grupo principal de Clientes:** Variable categórica que indica la segmentación de clientes, la cual dice que tipo de cliente es según su comportamiento de compra, o negocio. Se clasifican en Comercio de Alimentos, Comercio de no alimentos, Procesador de Alimentos, Prestador de Servicios, Compradores Individuales, Clientes de grandes Volúmenes, y Estaciones de Servicios.

- **Grupo de Cliente:** Variable categórica que indica la subdivisión de cada uno de los grupos principales de clientes, es decir, según el tipo de negocio que tiene, por ejemplo, restaurantes, hoteles, tenderos, etc.
- **División:** Variable categórica que indica el área en la cual se clasifican cada uno de los siguientes artículos. Makro cuenta con 3 áreas principales: Dry Food, Non Food y Fresh Food.
- **Departamento:** Variable categórica que indica s la subdivisión de cada una de las áreas que conforman la compañía. Su comportamiento detallado se muestra en la tabla 1 en la sección de anexos.
- **Articulo**: Son cada uno de los SKU's que conforman el portafolio que oferta Makro.
- **Venta:** Variable numérica que nos indica la venta en pesos colombianos de la compañía.
- **Lucro:** Variable numérica que nos indica la ganancia en pesos colombianos de la compañía.
- **Margen:** Variable numérica que nos indica la ganancia en porcentaje.
- **Facturas:** Variable numérica que nos indica el número de facturas registradas por cliente en los ultrimos 7 meses de la compañía.
- **Cantidad vendida:** Variable numérica que nos indica la cantidad de artículos vendida por cliente en los últimos 7 meses.
- **Clientes activos:** Variable numérica que nos indica el número de clientes que realizaron una compra de cualquier monto en los últimos 7 meses.

PROPUESTA METODOLÓGICA

Se define un algoritmo/técnica que se planea utilizar y se plantea qué otros algoritmos/técnicas son candidatos a utilizarse para solucionar la pregunta de interés. Se argumenta por qué el elegido es el adecuado dado el contexto del problema y de la organización. [15 puntos]

Existen dos características clave de los datos a tener en cuenta para tomar la decisión del algoritmo/técnica a emplear:

1. Gran volumen de datos (superior a 160.000)
2. Alta dimensionalidad

3. Presencia de outliers

PCA – Reducción de dimensionalidad
Dado que nuestros datos tienen alta dimensionalidad se revisará dentro del proceso de implementación la conveniencia de implementar la reducción de dimensionalidad a través de PCA.

Utilizando PCA entonces buscaremos simplificar la complejidad de nuestra base conservando solo aquellos componentes que nos conserven la mayor cantidad de información. Encontraremos así una cantidad de factores subyacentes que expliquen porcentajes superiores al 90% de la variabilidad de los datos. Calibraremos la cantidad de componentes a utilizar buscando encontrar mejores resultados.

Alternativa 1 – CLARA

Dado que el algoritmo de k-medias es altamente sensible a los outliers, podríamos evaluar en paralelo un modelo de k-medoides vs un modelo de k-medias que nos permita analizar el desempeño de ambos algoritmos. Dada la gran cantidad de datos, debemos recurrir al método CLARA (Clustering For Large Applications) el cual nos permite manejar datos de gran volumen y reducir el tiempo computacional, a través de muestreo.

De esta manera, en lugar de encontrar medoides para todo el conjunto de datos, encontraremos una pequeña muestra de los datos con un tamaño fijo para generar un conjunto óptimo de medoides en la muestra.

Como la pregunta pretende encontrar aquellos productos sobre los cuales la compañía maximizará sus ganancias, el agrupamiento en clústeres resulta útil porque nos permitirá agrupar los productos de acuerdo con su patrón de venta.

Se pretende entonces segmentar la base de acuerdo con los productos que presenten patrones de venta similares. Estos patrones de venta se establecerán basados en la información que se tiene sobre las compras de cada cliente en los días en los cuales se habilitan las promociones. Se establecerán los grupos de productos, para posteriormente definir cuáles de estos grupos son los de mayor margen.
Alternativa 2 – DBSCAN

Dado que al final lo que se pretende es incrementar las ventas buscando hacer las ofertas adecuadas a cada tipo de cliente, podríamos utilizar un algoritmo que nos permita a partir de un conjunto de datos históricos, predecir y recomendar a los clientes los productos relevantes en los cuales podría estar interesado.

De esta manera, se establecerían los grupos de productos de acuerdo a los tipos de clientes que visitan la tienda en cada uno de los días de promociones específicos para orientar las promociones y estrategias comerciales hacia los productos más probables a seleccionar por estos clientes.

DBSCAN nos permitirá entonces identificar clientes con patrones de compra similares para establecer cuáles de estos grupos tienen mayor margen y de esta manera, focalizar las estrategias a los productos que compra habitualmente este grupo de clientes.

Ventajas de seleccionar este algoritmo:
- Permite encontrar clústeres con formas y tamaños distintos
- Es muy robusto a outliers e incluso nos permitiría ubicarlos e identificarlos dentro de los datos

Desventajas:
- Podría tener un costo computacional alto

- La definición de épsilon y la métrica de distancia requiere un profundo conocimiento de los datos. Se recomienda calibrar los parámetros para encontrar los resultados que mejor se ajusten al problema a solucionar.

Bibliografía

Se citan los artículos mencionados en el texto, usando de forma correcta y consistente el estilo de referencia que se hay escogido usar (Chicago, APA, MLA, etc.). [5 puntos] Mariana

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

Anexos

Tabla 1.

Departamento	Ventas	Lucro	Margen	Facturas	Cantidad Vendida	Clientes Activos
Sugar/Coffee/Tea	$26,146,562,892	2,404,190,503	9.20%	330,892	797,306	201,413
Grains	$16,817,050,319	970,156,994	5.77%	239,228	845,971	139,414
Dairy / Others	$8,875,969,669	1,337,292,306	15.07%	445,922	660,737	267,254
Beers	$8,716,194,701	185,737,364	2.13%	87,973	1,894,801	47,303
Oils	$8,319,937,396	723,839,593	8.70%	80,083	162,746	45,628
Candys / Snacks	$7,236,911,577	1,793,328,378	24.78%	668,220	971,677	372,126
Personal Care	$5,738,380,252	661,365,734	11.53%	332,431	404,943	204,096
Hygiene	$5,560,748,686	625,691,091	11.25%	235,890	511,569	141,337
Wine / Spirits	$5,431,995,028	308,956,324	5.69%	111,650	383,653	72,951
Condiments	$5,392,934,407	911,348,054	16.90%	406,487	683,099	248,101
Milk / Cereals	$5,191,798,563	761,820,746	14.67%	148,743	359,227	87,558
Flours	$4,559,497,954	515,146,091	11.30%	222,383	549,737	132,065
Laundry	$4,355,379,165	280,965,188	6.45%	160,183	944,172	95,944
Soft Drinks	$4,091,097,033	48,345,630	1.18%	121,905	662,681	61,704
Milk	$3,457,940,390	423,363,429	12.24%	100,542	181,311	55,545
Pork / Others	$3,391,678,319	296,312,790	8.74%	59,307	243,885	35,131
Beef	$3,371,019,401	150,723,831	4.47%	54,106	138,293	34,423
Cold Meat	$3,053,559,020	523,899,553	17.16%	172,695	236,009	100,692
Canned / Preserves	$2,992,605,861	359,750,630	12.02%	145,830	292,139	88,482
Fruver	$2,932,041,961	188,016,477	6.41%	480,014	741,454	267,222
Frozen	$2,896,133,681	632,810,190	21.85%	132,517	209,770	81,484
Cleaning	$2,841,150,117	482,987,882	17.00%	174,615	244,063	111,807
Poultry	$2,776,495,086	315,959,992	11.38%	86,956	223,756	48,937

Bakery	$2,614,042,460	488,268,353	18.68%	296,659	437,740	168,355
Disposible	$2,317,087,594	501,148,799	21.63%	105,042	271,189	69,634
Water / Juices	$2,284,571,830	269,017,298	11.78%	195,007	737,366	100,785
Cleaning/Organiz	$2,113,129,454	594,615,540	28.14%	177,306	243,623	107,274
Electro	$2,020,494,323	289,750,666	14.34%	4,147	4,106	3,582
Seafood	$1,851,478,750	285,353,856	15.41%	57,408	81,659	36,811
Liquid	$1,816,321,040	70,055,915	3.86%	21,765	207,000	59
Eggs	$1,598,801,434	43,846,008	2.74%	56,977	111,716	29,035
Textil	$1,165,981,302	355,533,603	30.49%	37,279	57,767	25,812
Pet	$1,135,036,592	234,002,191	20.62%	43,158	54,217	27,982
Cooking Utilities	$1,059,504,975	295,135,896	27.86%	31,222	39,630	24,028
Office	$713,624,797	143,903,839	20.17%	17,262	58,330	12,921
Automotive	$689,720,766	124,510,210	18.05%	6,923	9,034	5,104
Tabaco	$607,443,456	12,738,564	2.10%	9,574	292,874	3,116
Fresh Bakery	$423,956,532	147,439,381	34.78%	64,822	143,737	27,900
Table Services	$344,259,099	99,059,275	28.77%	12,522	50,472	9,578
Furniture	$257,801,868	61,814,192	23.98%	1,447	3,087	1,202
Diy	$175,037,614	44,513,877	25.43%	13,659	19,310	9,380
Seasonal	$78,365,281	(21,754,957)	-27.76%	3,104	3,534	1,950
Service	$69,849,597	24,112,083	34.52%	27,609	44,607,149	3,408