FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD DE BUENOS AIRES

Ciencia de Datos para la Toma de Decisiones

Trabajo de Aplicación

Integrantes:

Iván Benjamín Gancedo	108794
Andrés Mocnik	108875

Segmentación para empresa de cosmética

Introducción

La empresa bajo análisis opera predominantemente a través de un modelo de venta directa, sustentado por una extensa red de consultoras y revendedores independientes. Estos emprendedores son el principal canal de comercialización y relación con el cliente final, gestionando sus ventas mediante catálogos y pedidos por ciclo. Adicionalmente, la compañía ha incorporado un canal de venta online directo al consumidor, que actualmente representa un porcentaje minoritario (aproximadamente el 5%) de sus ingresos totales, complementando así la estrategia de venta principal y ofreciendo una alternativa de compra digital. Este estudio se centrará en analizar datos generados a partir de la actividad de su red de revendedores.

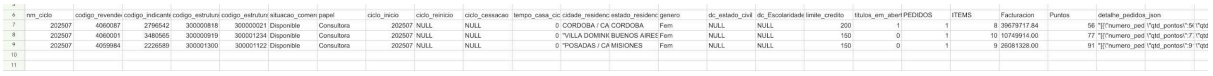
En un mercado competitivo, comprender a fondo tanto la fuerza de ventas como las preferencias locales de los clientes es fundamental. La segmentación de revendedores permitirá a la empresa enfocar sus recursos de manera más eficiente, ofreciendo apoyo y motivación adecuados a cada perfil.

Objetivo:

El objetivo es segmentar las canastas de productos de los revendedores para identificar patrones de compra comunes y tipos de canastas predominantes. Esto permitirá entender mejor las preferencias de los revendedores, optimizar el surtido de productos, personalizar ofertas y mejorar estrategias de marketing y ventas.

Descripción del dataset:

El dataset para este análisis proviene de una base histórica de pedidos y logística de la empresa, comprendiendo aproximadamente 50,000 pedidos únicos distribuidos entre unos 30,000 revendedores activos. La información se organiza en una estructura tabular con 53 variables, que abarcan desde datos sociodemográficos del revendedor (género, ciudad, escolaridad), su información comercial (ciclo de alta, estructura, situación crediticia), hasta resúmenes del historial de pedidos (cantidad de ítems, valor, puntos) y detalles logísticos (transportadora, cajas). Si bien existe un porcentaje menor de datos faltantes (2-5%) en algunas variables sociodemográficas, la integridad de los datos de pedidos es alta.



	nm_ciclo	codigo_revende	codigo_indicante	codigo_estruturi	codigo_estruturi	situacao_comen	papel	ciclo_inicio	ciclo_reinicio	ciclo_cessacao	tempo_casa_cic	cidade_residenc	estado_residenc	genero	dc_estado_civil	dc_Escolaridade	limite_credito	ttulos_em_aber	PEDIDOS	ITEMS	Facturacion	Puntos	detalhe_pedidos_json
7	202507	4060087	2796542	300000818	300000021	Disponible	Consultora	202507	NULL	NULL	0	CORDOBA / CA	CORDOBA	Fem	NULL	NULL	200	1	1	8	3967977.84	56	"[{""numero_ped	\"qtd_pontos\":5(	\"qtd
8	202507	4060001	3480565	300000919	300001234	Disponible	Consultora	202507	NULL	NULL	0	"VILLA DOMINI	BUENOS AIRES	Fem	NULL	NULL	150	0	1	10	1074914.00	77	"[{""numero_ped	\"qtd_pontos\":7,(	\"qtd
9	202507	4059984	2226589	300001300	300001122	Disponible	Consultora	202507	NULL	NULL	0	"POSADAS / CA	MISIONES	Fem	NULL	NULL	150	0	1	9	26081328.00	91	"[{""numero_ped	\"qtd_pontos\":9	\"qtd
10																							
11																							

Para el objetivo específico de **segmentar las canastas de productos**, el componente más crucial de este dataset reside en una columna que contiene información detallada de los pedidos en formato **JSON**. Aunque el dataset principal tiene una fila por revendedor y ciclo con métricas agregadas, este campo JSON anidado es el que revela la composición granular de cada transacción individual.

Dentro de este campo, la estructura es jerárquica y contiene, por cada pedido:

- **Información agregada del pedido:** como cantidad de puntos y cantidad de ítems totales del pedido, valor total, y fechas.
- **Subestructura ítems:** Esta es una **colección (lista/array) fundamental** para nuestro análisis. Cada elemento de esta lista representa un producto individual dentro del pedido e incluye:
    - Código de producto: El identificador único del producto, que será la pieza central para determinar qué artículos componen cada canasta.
    - Tipo de producto: Una categorización del producto, que puede ser usada para un análisis a un nivel más agregado o para enriquecer la interpretación de los segmentos de canastas.
    - Cantidad de ítems: La cantidad de ese producto específico, permitiendo análisis que consideren no solo la presencia sino también el volumen de cada producto en la canasta.
    - Otros campos como “Valor total a pagar” y “Total de puntos” a nivel de ítem, que podrían usarse secundariamente para caracterizar las canastas.
- **Subestructura logística:** Aunque contiene detalles como transportadora y cajas, esta parte del JSON es menos relevante para la segmentación de la *composición* de la canasta de productos en sí.

Métodos analíticos a emplear: libro tibshirani, y algo MLS

Para segmentar las canastas, nos centraremos en el contenido del JSON de cada pedido, específicamente en los códigos de producto o tipo de producto dentro de cada pedido.

Primero se hará un clustering de pedidos, cuyo objetivo es agrupar los pedidos que son similares entre sí basándose en los productos que contienen. Cada grupo resultante (clúster) representará un "tipo de canasta" o un patrón de compra común.

Para resolver el problema, se usará una matriz pedido-producto con los datos del JSON, usando la frecuencia de cada producto, del siguiente modo:

Id_Pedido	ProdA	ProdB	ProdC	…
1	2	0	1	…
2	0	0	0	…
…	…	…	…	…

Luego se haría un preprocesamiento de los datos, donde reduciremos la dimensionalidad mediante PCA (Análisis de Componentes Principales). El PCA transformará las columnas de productos en un número menor de "componentes principales" que capturan la mayor parte de la varianza. El clustering se haría sobre estos componentes.

En lo que respecta al clustering, se usará un **Hierarchical K-Means**, que consiste en primero hacer un clustering jerárquico sobre una muestra parcial para tener una base para luego aplicar K-Means sobre el total de los datos. Esto debería dar un mejor número de clusters.

Finalmente y como análisis complementario se usará un **Análisis de Canasta de Mercado** (Market Basket Analysis - MBA) que emplea algoritmos como Apriori o FP-Growth con el objetivo primordial de descubrir reglas de asociación entre productos. Esto significa identificar patrones del tipo "si un revendedor compra el Producto A, existe una alta probabilidad de que también adquiera el Producto B". Para cuantificar estas relaciones, se utilizan métricas clave:

- El **Soporte** indica la frecuencia con la que un conjunto de ítems aparece conjuntamente en todas las transacciones;
- La **Confianza** mide la probabilidad de comprar el Ítem B dado que ya se ha comprado el Ítem A;
- El **Lift** revela cuánto más probable es comprar el Ítem B si se compra el Ítem A, en comparación con la probabilidad general de comprar el Ítem B, indicando la fuerza de la asociación más allá de la popularidad individual de los productos.

La utilidad principal de este análisis radica en identificar productos que frecuentemente se venden juntos, lo cual es invaluable para diseñar promociones cruzadas efectivas, crear paquetes de productos atractivos o mejorar la disposición del catálogo para fomentar compras adicionales.