Estudiante: Juan Camilo Vergara Arenas
Proyecto de Grado: “Modelo de recomendación de portafolio óptimo para aumentar participación en mostradores”
Universidad Icesi

Introducción

En el ámbito comercial, el área de Ventas enfrenta constantemente el desafío de comprender de manera integral la dinámica de sus diversos canales. Este análisis resulta fundamental para el diseño de estrategias efectivas que permitan posicionar adecuadamente los portafolios de productos y maximizar su impacto en cada canal de ventas.

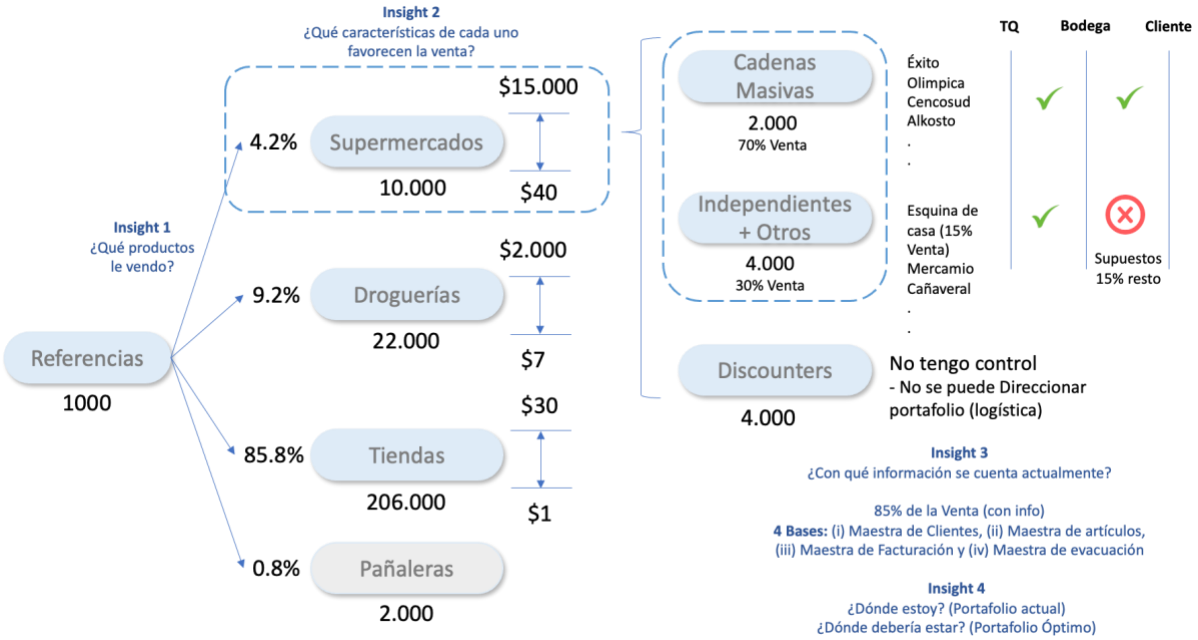
Entre estos canales, el autoservicio destaca como un segmento clave en las estrategias comerciales. Este tipo de canal abarca puntos de venta donde los consumidores pueden seleccionar y adquirir productos directamente, ofreciendo una experiencia de compra ágil y autónoma. A su vez, el canal de autoservicios se clasifica comúnmente en tres categorías:

- **Grandes cadenas:** Supermercados y tiendas de alcance nacional o regional, caracterizados por una amplia oferta de productos.
- **Independientes:** Negocios autónomos con una oferta más limitada, pero que mantienen una conexión directa con los consumidores locales.
- **Discounters:** Tiendas enfocadas en ofrecer precios competitivos/más bajos mediante estrategias de optimización de costos y una selección básica de productos.

Este proyecto se centra en el análisis de dos de estas categorías (Grandes Cadenas e Independientes) con el propósito de identificar dinámicas específicas que permitan desarrollar estrategias de venta, adaptadas a las particularidades de cada cliente (Acciones en pos del entendimiento del comportamiento del canal de autoservicios).

Contexto y antecedentes

Como marco general, el proyecto nace en el área de Ventas de la compañía Tecnoquímicas, debido a un pedido corporativo del CEO que busca (i) entender sus principales canales/clientes, con el propósito de (ii) identificar oportunidades que permitan llegar con el portafolio adecuado de productos en pos de (iii) alinear correctamente pedidos de negocios y (iv) generar mayor competitividad.



El proyecto se centraliza en un modelo que tiene definido unos momentos o estadios:

1. **Momento Inicial:** todo empieza con el portafolio de productos de la empresa de cada uno de los negocios (17) y la necesidad de identificar cuál es el mix del portafolio que se venden a los distintos tipos de clientes. El portafolio de cada negocio tiene unas características y unos pedidos solicitados por el negocio para sembrar y distribuir en los distintos canales y tipos de clientes.
   La pregunta que se debe resolver aquí es ¿Cuáles son los productos que se le venden a los distintos clientes?

   Para ello se cuenta con bases como: Maestra de artículos, donde se puede explorar la información relacionada con los productos, como sus marcas, líneas, moléculas, terapias/audiencias, etc

2. **Momento en la Distribución:** actualmente, existen distintos tipos de clientes a los que tecnoquímicas les vende con el propósito de llegar a sus consumidores finales. Estos se dividen en 4 grandes categorías (conceptualmente): (i) Supermercados, (ii) Droguerías, (iii) Tiendas y (iv) Pañaleras.

   - Los supermercados son las grandes cadenas de mercados que manejan un gran volumen de productos y transacciones diarias (la expectativa acá es volumen en portafolio),
   - Las droguerías son clientes especializados que requieren un enfoque personalizado (portafolio adaptado) debido a la naturaleza regulada de su industria
   - Las tiendas son clientes versátiles que abarcan desde pequeños comercios hasta cadenas más grandes. Su enfoque se basa en la diferenciación y el servicio al cliente
   - Pañaleras son un nicho especializado en productos para bebés y niños

En este punto el foco del proyecto está en los supermercados ya que son un grupo de clientes que tienen un gran volumen de ventas a pesar de ser menor cantidad como número de clientes, lo que los hace muy representativos.

Dentro de esta categoría hay unas subcategorías importantes: (i) cadenas masivas, (ii) independientes y (iii) discounters

- Las cadenas masivas hacen referencia a estos supermercados realmente grandes, que manejan alto volumen y que tienen incluso múltiples puntos de venta en distintos lugares/regiones, por ejemplo: el grupo éxito, alkosto, Olímpica
- Los independientes son supermercados medianos con un manejo de volumen medio y que cuentan también con múltiples puntos de venta en distintos lugares/regiones, por ejemplo: Cañaveral, Esquina de la casa, Mercamio
- Los discounters son tiendas minoristas que se especializan en ofrecer productos a precios reducidos, a menudo a expensas de un servicio al cliente más personalizado y un ambiente de compra más limitado

Aquí el interés es responderse ¿Cuáles son las características de estos clientes que podrían favorecer la venta?

Para ello se cuenta con bases como: Maestra de clientes, donde se puede explorar la información relacionada con los productos, como sus marcas, líneas, moléculas, terapias/audiencias, etc

Objetivos

Pregunta Problema

¿Cuál es el portafolio óptimo (Mix de marcas de productos ideal) para cada cliente/mostrador de Autoservicios de TQ en pos de aumentar la participación en los mostradores?

Portafolio Óptimo: Hace referencia a la combinación de cantidades de Marcas que se deben recomendar por cada Trimestre en cada punto de venta

Objetivo General

Diseñar un modelo que permita proponer un mix de productos óptimo para llegarle a cada cliente de autoservicios en pos de aumentar la participación en los mostradores

Objetivos específicos

1. Definir el dataset requerido para el modelamiento a partir de las consultas necesarias en el sistema adecuado (Delta)

2. Realizar el análisis exploratorio de los datos para (i) entender la información a la luz de los negocios y (ii) organizar la data relevante para modelamiento

3. Diseñar un modelo de clasificación/recomendación que estime la demanda de productos para cada cliente, optimizando el mix propuesto y aumentando la participación en los mostradores a partir de aprendizaje no supervisado

Metodología

Identifica, justifica y describe la metodología utilizada para el desarrollo del proyecto.

La metodología consiste en:

1. Entendimiento del negocio
2. Entendimiento de datos / Preparación de los datos
    - Modelo relacional de la data
    - Organización de la data
    - Análisis de datos
3. Modelamiento
4. Valoración del modelo

Esta se apoya en el modelo CRISPDM, el cual se compone en esencia de 7 fases: **(i) comprensión del negocio**, **(ii) comprensión de los datos**, **(iii) preparación de los datos**, **(iv) Modelamiento**, (v) evaluación, (vi) despliegue y (vii) revisión y mantenimiento. De estas, considero que relevantes para el proyecto en PDGI serían las primeras 4 fases

Marco Teórico

Negocio:

Conjunto de actividades, estrategias y procesos que permiten a una organización generar valor mediante la comercialización de productos o servicios. En este contexto, el negocio se estructura en torno a portafolios de productos diseñados para satisfacer las necesidades de mercados específicos.

Línea:

Categoría o familia de productos dentro de un portafolio, agrupados por sus características o funcionalidad compartida. Las líneas permiten organizar y segmentar el portafolio para atender diferentes necesidades del consumidor.

Marca:

Identidad comercial de un grupo de productos dentro de una línea que los distingue en el mercado. La marca abarca valores, percepciones y atributos asociados que buscan atraer y fidelizar a los clientes.

Artículo:

Producto específico dentro de una marca, definido por atributos como tamaño, sabor, presentación o empaque. Los artículos representan el nivel más granular dentro de la estructura del portafolio.

Canal de Venta:

Medio a través del cual los productos son distribuidos y vendidos al consumidor final. Los canales de venta pueden incluir autoservicios, tiendas de conveniencia, mayoristas, e-commerce, entre otros, adaptándose a las dinámicas de compra de cada segmento de clientes.

Venta Colocación:

Volumen o valor de productos entregados por la empresa a sus clientes directos (por ejemplo, cadenas de autoservicio o distribuidores). Representa la primera etapa en la cadena de distribución, donde la empresa asegura la disponibilidad del producto en los puntos de venta.

Venta Evacuación:

Volumen o valor de productos vendidos por los clientes directos de la empresa al consumidor final. Esta métrica refleja el desempeño del producto en el mercado y la aceptación por parte de los consumidores.

Vectorización: Proceso donde se define para cada artículo en cuál punto de venta debe estar

Subcanal (Familia estadística): Agrupación de clientes homogéneos, según su forma de atención al consumidor final, para su manejo estratégico comercial. Ej.: Tiendas, Distribuidores, Droguerías Independientes, Autoservicios en Cadena, Entidades Públicas, Almacén Agroveterinario, entre otras.

Unidades estadísticas: Medida estándar de volumen que permite la agregación y/o comparación de acuerdo con un criterio seleccionado de segregación especifico del

producto. Estas unidades pueden estar expresadas en dosis, toneladas, kilos, yardas, etc.

Siembra: Posicionamiento de un producto nuevo en los puntos de venta.

Procesos estratégicos: Procesos diseñados para medir, monitorear y controlar las actividades del negocio. Estos procesos aseguran que los misionales y los de apoyo se diseñen y ejecuten de forma que logren las metas operacionales, financieras, regulatorias y legales.

Principio activo: Compuesto o mezcla de compuestos que tiene una acción farmacológica.

Precio base: Valor base o precio de lista de un artículo para clientes comerciales establecido por la unidad de negocio con base en el estudio de precios, al cual se aplican los descuentos correspondientes al canal o al cliente para lograr la posición relativa de precios en el mercado.

Penetración: Porcentaje de los clientes manejantes de un artículo de la compañía sobre el total de los clientes manejantes de la categoría.

Marca: Nombre que identifica una agrupación de artículos que comparten una misma imagen construida hacia el consumidor.

Línea: Agrupación interna de productos con características comunes, enmarcados en la estrategia de mercadeo para su manejo administrativo y comercial.

Forma farmacéutica: Forma física que caracteriza al producto farmacéutico terminado, comprimidos, cápsulas, jarabes, supositorios, etc.

Distribución: Presencia de los artículos en cada uno de los establecimientos comerciales manejantes de las categorías de TQ.

Colocación: Unidades o pesos que se le facturan al cliente.

Canal: Medio a través del cual los fabricantes o distribuidores ponen a disposición de los consumidores los productos para que los adquieran.

Artículo: Bien manufacturado y/o comercializado que se ofrece a un mercado para su adquisición, uso y consumo.

Actividad (Ventas): Diseño basado en diagnósticos precisos, encaminados a asegurar que el acto de compra se cierre en el punto de venta. Estos diseños están enmarcados en las estrategias definidas en el plan de canal y en el plan de marca para el canal.

Estado del arte

Revisión de artículos

1. "**A Data-Driven Approach to Product Assortment Optimization in Retail"**

Este artículo se focaliza en el uso de métodos basados en datos en las operaciones de minoristas, destacando cómo la digitalización y el acceso a grandes volúmenes de datos han transformado la investigación y práctica en este campo.

Por un lado, se explicita el cambio de la gestión de la operación en este tipo de negocios tipo retail. donde se pasa de estar centralizados en sistemas de producción a un concepto más amplio de gestión de la cadena de valor completa a través del uso fundamentado de los datos. Ha crecido el uso de técnicas de análisis de datos donde el nuevo enfoque es utilizar grandes volúmenes de datos para tomar decisiones de manera informada.

Y, particularmente se menciona la gestión de inventarios en pos de garantizar disponibilidad de productos.

El artículo aborda varios temas:

- La **optimización del asortimiento**: muestran la importancia que tiene para las operaciones de minoristas y la gestión de ingresos. Se centra en la selección del conjunto óptimo de productos que un minorista debe ofrecer a los clientes, con el objetivo de maximizar los ingresos y la satisfacción del cliente. Esto presenta un problema combinatorio, ya que no siempre es viable o deseable ofrecer todos los productos disponibles.

  Esto es en esencia lo que busca mi proyecto para tecnoquímicas y sus clientes de autoservicios.

  En ese sentido, se establecieron unos modelos para:

  Modelos de Elección Discreta:

- **Multinomial Logit (MNL)**: Modela las elecciones de los clientes entre varios productos (o la opción de no comprar) basándose en la utilidad percibida, que incluye un término de ruido aleatorio.
- **Mixed Multinomial Logit (MMNL)**: Extiende el MNL al permitir que las utilidades medias sean variables aleatorias, lo que puede representar diferentes tipos de clientes.
- **Nested Logit (NL)**: Permite que los clientes elijan primero un "nido" (grupo de productos) y luego un producto específico dentro de ese nido.

Esta sección en el proyecto aporta de la siguiente manera:

Selección de Productos: Se podrían implementar los enfoques descritos para determinar el conjunto óptimo de productos (mix) que permitiría maximizar las ventas en los mostradores. Es decir, usar modelos como MNL o MMNL para comprender cómo diferentes grupos de clientes valoran los productos.

Modelado de Preferencias del Cliente: Utilizar modelos de elección discreta para segmentar a tus clientes en función de sus preferencias y comportamientos de compra

- **Assortment óptimo:** Una vez que se estiman los parámetros de los modelos de elección, se puede determinar el assortment óptimo resolviendo un problema de optimización.
    - **assortment estático**, se busca el assortment óptimo que maximiza el beneficio esperado en un único intento. Si se utilizan modelos MNL (Multinomial Logit) con parámetros de utilidad conocidos, el problema se puede resolver eficientemente considerando solo assortments ordenados por ingresos.

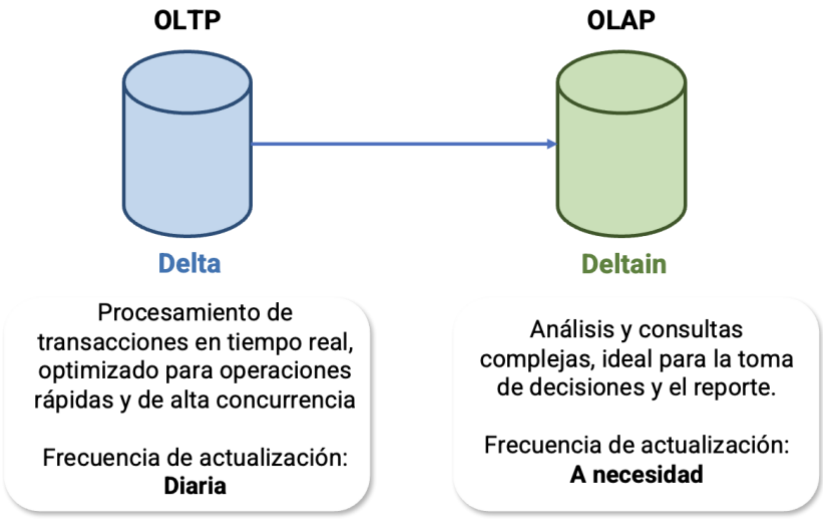
Resultados

1. Entendimiento de la data

.

El área cuenta con información para poder crear un modelo que permita estimar/recomendar el mejor mix de productos para cada tipo de cliente. Esto, a

través de las bases de datos que se han construido en el tiempo y que se encuentran organizadas bajo un esquema típico de arquitectura de datos:



Por un lado se cuenta con un sistema transaccional llamado Delta que recopila información relacionada con facturación, demografía, precios, evacuación, etc que se capturan en tiempo real y se actualizan con una frecuencia diaria que permite tener un data warehouse actualizado con información relacional de cada acción de venta con los clientes

Y, por otro lado, se cuenta con un sistema de visualización que permite realizar consultas a esta data warehouse, llamado Deltain. Desde la plataforma es posible realizar un ajuste de los informes de acuerdo a la selección de variables de interés y de esta manera es posible generar las bases de datos requeridas para el desarrollo del proyecto

Aquí el interés es responderse ¿Dónde estoy (portafolio actual) en términos de variables críticas de venta? y ¿ Dónde debería estar (portafolio óptimo) en términos de variables críticas de venta?

De esta manera, se consultaron 3 bases grandes (1 para cada tipo de cliente de interés: **91** Droguerías independientes, **93** Autos Independientes y **94** Autos en cadenas). Y, se realizó un proceso de pretratamiento de la data bajo ciertos parámetros establecidos con el experto:

1. Parámetro de tiempo: Se estableció que la data tendría una historia de año móvil entre 2023 y 2024 y que el manejo del Dataset a modelar sería organizado por trimestres. En ese sentido las fechas de los 3 dataset iria de octubre 2023 a septiembre 2024

2. Parámetro portafolio de productos: Se estableció que la data debía organizarse por Negocio, Línea y Marca de producto con una consideración especial para el Negocio: “UN Absorbentes” debido a que su Marca no tiene una información tan diciente/completa entonces se debe concatenar la línea con la marca para tener la “Marca” definitiva

3. Parámetro de clientes: Se estableció primero los tipos de clientes de interés: (i) Cadenas grandes (94), Independientes (93) y una droguería de comportamiento de supermercado: Comfandi (91).

4. Parámetro de Venta: Se estableció que la venta de colocación presente en el subcanal 93 es suficientemente diciente para explicar el comportamiento en el punto de venta y esto es porque al tratarse de clientes pequeños la relación de cliente padre y cliente punto de venta es de 1 en su mayoría y esto quiere decir que lo que se le vende al cliente es lo que se vende directamente al punto de venta

Sin embargo, por otro lado para el subcanal 94 y 91 se tienen clientes más grandes y en ese sentido la relación de cliente padre y cliente punto de venta es distinta de 1 y es una relación muy grande, lo que indica que la distribución de la venta de colocación en cada punto de venta es diferente y es información que toca extraer de la venta de evacuación

5. Parámetros adicionales calculados: De manera distinta se requiere información que no se puede descargar directamente de las bases sino que se debe calcular para enriquecer el modelo final. Esta información es la siguiente:

- Frecuencia: Es un indicador que mide cuántas veces un cliente realiza compras de un producto en un período específico, como por ejemplo un trimestre. Esto permite entender el comportamiento de compra de los clientes y clasificar su nivel de actividad

- Rotación: Es un indicador que mide la cantidad promedio de unidades que un cliente compra de un producto cada vez que realiza una compra. Este indicador complementa la frecuencia de compra y permite analizar la intensidad del consumo

Unidades Comerciales / Frecuencia de compra

De estas se tiene el siguiente listado de variables de interés:

Variables Transversales	Descripción
Pais (ID)	País (En este caso solo se acotó a Colombia, entonces se filtró la consulta únicamente para Colombia)
Cod_Negocio	Código de la unidad de negocio de la compañía
Negocio_Nombre	Nombre de la Unidad de negocio de la compañía
Mes	Mes de la venta
Cod_Subcanal	Hace referencia al tipo de canal para llegar (Eq al tipo de cliente) Ej: Autoservicios independientes
Subcanal_Nombre	Nombre del canal
Cod_Clipadre	Código del Cliente Padre
Clipadre_Nombre	Nombre del Cliente Padre (es el nombre del cliente que actúa como sombrilla de cada cliente. Ejemplo Grupo Éxito)
Cod_Cliente	Código del Cliente
Cliente_Nombre	Nombre del Cliente (Tipo de venta/Bodega)

Variables Colocación	Descripción
Venta + Dscto	Valor de venta de producto
Unidades vta comercial	Unidades de artículos vendidos

Variables Evacuación	Descripción
Evac_VNETA1	Valor de venta de producto
UComerciales	Unidades de artículos vendidos

2. Pretratamiento de la data

El pretratamiento de los datos representó un desafío clave e importante, ya que implicó la creación de uno o varios datasets necesarios para desarrollar distintos modelos enfocados en recomendar un mix de productos personalizado para cada cliente del canal de autoservicios.

En este contexto, una parte significativa del proyecto se centró en la construcción de estos datasets, etapa fundamental previa al modelado.

Este proceso se realizó en dos fases:

2.1. Análisis Exploratorio de Datos

Para garantizar que la información consultada del sistema cumplía con la estructura, coherencia y validez necesaria para modelar, entonces se realizó lo siguiente:

1. **Revisión de nombres:** Se vió la estructura general de la data y se realizaron los ajustes pertinentes para que la información tuviera identificada el tipo de información que tenía de manera coherente con el negocio. A modo de ejemplo, se puede ver como se encontraba la estructura inicial de la información:

```
Data columns (total 22 columns):
 #   Column         Non-Null Count    Dtype
---  ------         --------------    -----
 0   Pais (ID)      1048573 non-null  object
 1   Negocio        1048573 non-null  int64
 2   Unnamed: 2     1048573 non-null  object
 3   Linea          1048573 non-null  int64
 4   Unnamed: 4     1048573 non-null  object
 5   Marca          1048573 non-null  object
 6   Unnamed: 6     1048573 non-null  object
 7   Canal          1048573 non-null  object
 8   Subcanal       1048573 non-null  int64
 9   Unnamed: 9     1048573 non-null  object
 10  Clipadre       1048573 non-null  int64
 11  Unnamed: 11    1048573 non-null  object
 12  Unnamed: 12    1048573 non-null  int64
 13  Unnamed: 13    1048573 non-null  int64
 14  Cliente        1048573 non-null  object
 15  Unnamed: 15    991184 non-null   object
 16  Unnamed: 16    961149 non-null   float64
 17  Ano            1048573 non-null  int64
 18  Mes            1048573 non-null  object
 19  Indicadores    0 non-null        float64
 20  Evac_VNETA1    1048573 non-null  float64
 21  UComerciales   1048573 non-null  float64
dtypes: float64(4), int64(7), object(11)
```

   De esta manera se realizaron los ajustes necesarios para llegar a unas estructuras como las siguientes:

```
Pais (ID)
Cod_Negocio
Nombre_Negocio
Cod_Linea
Nombre_Linea
Cod_Marca
Nombre_Marca
Mes
Ano
Canal
Cod_Subcanal
Nombre_Subcanal
Cod_Clipadre
Nombre_Clipadre
Cod_Desconocido1
Cod_Desconocido2
Cod_Cliente
Nombre_Cliente
Cod_Desconocido3
Indicadores
Venta_Colocacion
Unidades_Comerciales_Colocacion
```

```
Pais (ID)
Cod_Negocio
Nombre_Negocio
Cod_Linea
Nombre_Linea
Cod_Marca
Nombre_Marca
Canal
Cod_Subcanal
Nombre_Subcanal
Cod_Clipadre
Nombre_Clipadre
Cod_Desconocido1
Cod_Desconocido2
Cod_Cliente
Nombre_Cliente
Cod_Desconocido3
Ano
Mes
Indicadores
Venta_Evacuacion
Unidades_Comerciales_Evacuacion
```

En este caso se manejó dos estructuras una para los dataset de la venta de colocación (Izquierda) y otro para la venta de evacuación (derecha). De estos ajustes habían códigos desconocidos que luego se validaron con el experto de negocio y no eran relevantes para el ejercicio así que se eliminaron

2. **Revisión de nulos:** En este paso se buscó identificar la presencia de datos nulos en las bases de datos y se definieron acciones para cada tipo de evento encontrado.

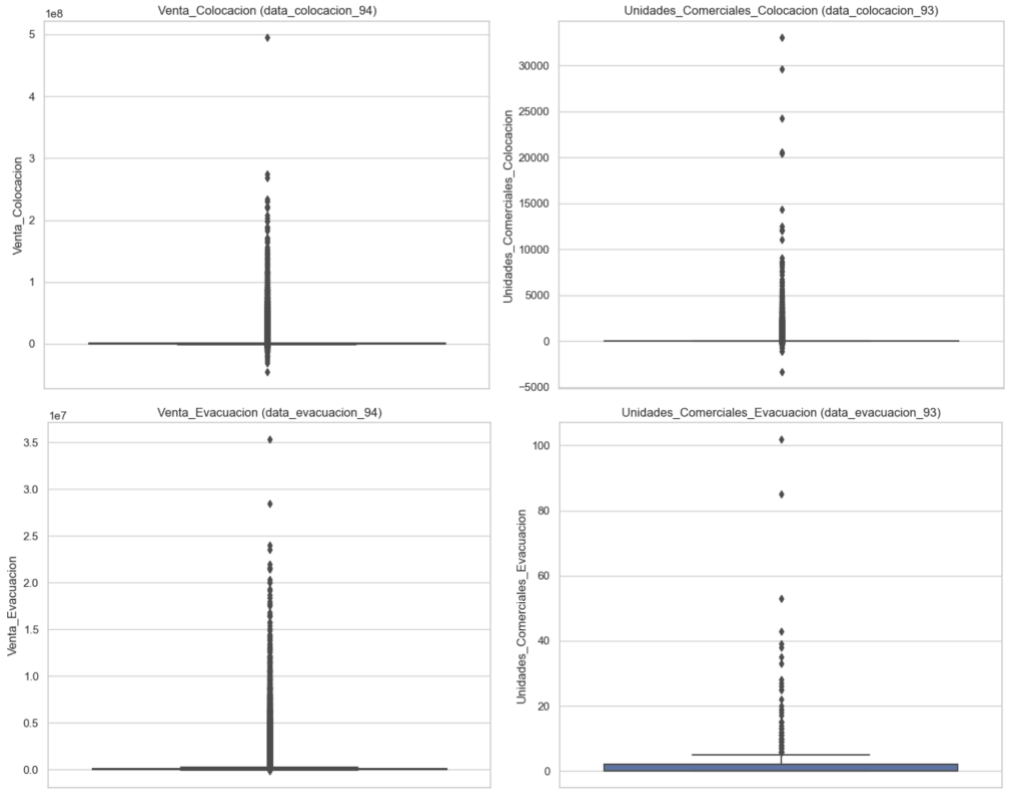
```
Valores nulos por columna:

Pais (ID)                  0
Division                   0
DivisionNombre             0
Negocio                    0
NegocioNombre              0
Articulo                   0
ArticuloNombre             0
ArticuloPresentacion       0
Mes                        0
Subcanal                   0
SubcanalNombre             0
Clipadre                   0
ClipadreNombre             0
Cliente                    0
ClienteNombre              0
Tipo Establecimiento       0
Venta + Dscto              0
Unidades vta comercial     0
dtype: int64
```

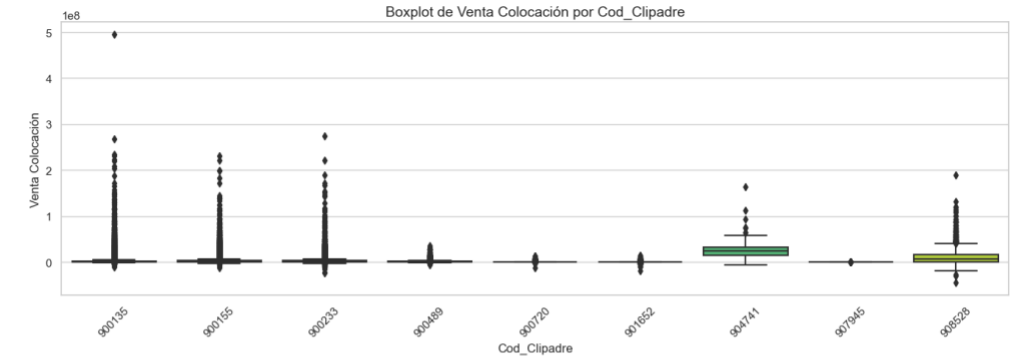
```
data_evacuacion_94:
Pais (ID)                         0
Cod_Negocio                       0
Nombre_Negocio                    0
Cod_Linea                         0
Nombre_Linea                      0
Cod_Marca                         0
Nombre_Marca                      0
Canal                             0
Cod_Subcanal                      0
Nombre_Subcanal                   0
Cod_Clipadre                      0
Nombre_Clipadre                   0
Cod_Desconocido1                  0
Cod_Desconocido2                  0
Cod_Cliente                       0
Nombre_Cliente                57389
Cod_Desconocido3              87424
Ano                               0
Mes                               0
Indicadores                 1048573
Venta_Evacuacion                  0
Unidades_Comerciales_Evacuacion   0
```

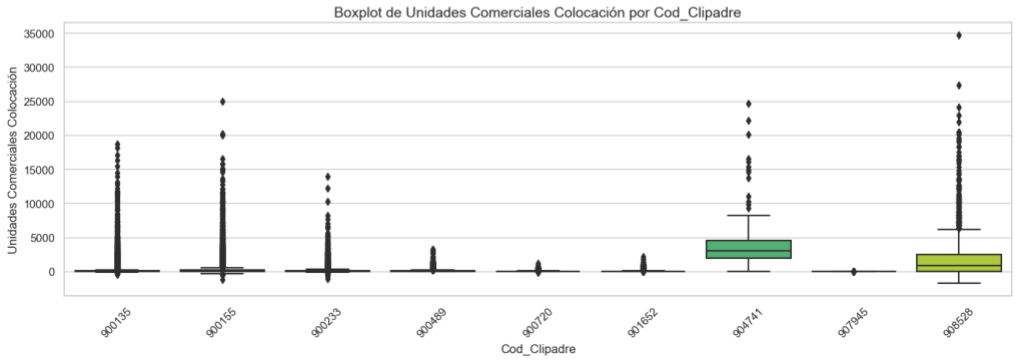
A la final se observó solo información nula en los nombres de clientes de la información de evacuación pero como se tenía la información de los códigos de estos clientes entonces no se optó por hacer nada con estos nulos porque se trabajaría con los códigos. De resto la información no contaba con nulos.

De igual manera, se decidió eliminar la información que no aportaba valor como los códigos desconocidos y los Indicadores

3. Atípicos:


Venta_Colocacion (data_colocacion_94)
1e8
Venta_Colocacion
Unidades_Comerciales_Colocacion (data_colocacion_93)
Unidades_Comerciales_Colocacion
Venta_Evacuacion (data_evacuacion_94)
1e7
Venta_Evacuacion
Unidades_Comerciales_Evacuacion (data_evacuacion_93)
Unidades_Comerciales_Evacuacion


Boxplot de Venta Colocación por Cod_Clipadre
1e8
Venta Colocación
900135
900155
900233
900489
900720
901652
904741
907945
908528
Cod_Clipadre


Boxplot de Unidades Comerciales Colocación por Cod_Clipadre
Unidades Comerciales Colocación
900135
900155
900233
900489
900720
901652
904741
907945
908528
Cod_Clipadre

	Venta_Colocación_94	Unidades_Colocación_94	Venta_Colocación_93	Unidades_Colocación_93
count	1.300700e+05	130070.000000	8.319310e+05	831931.000000
mean	1.539949e+06	143.065419	1.836008e+05	15.942493
std	6.904521e+06	749.388933	7.560338e+05	108.799101
min	-4.523713e+07	-1764.000000	-5.920039e+07	-3359.000000
25%	3.994190e+03	0.000000	7.980000e+03	0.000000
50%	1.511752e+05	12.000000	5.454500e+04	3.000000
75%	5.577270e+05	42.000000	1.584478e+05	12.000000
max	4.949273e+08	34680.000000	1.118887e+08	33024.000000

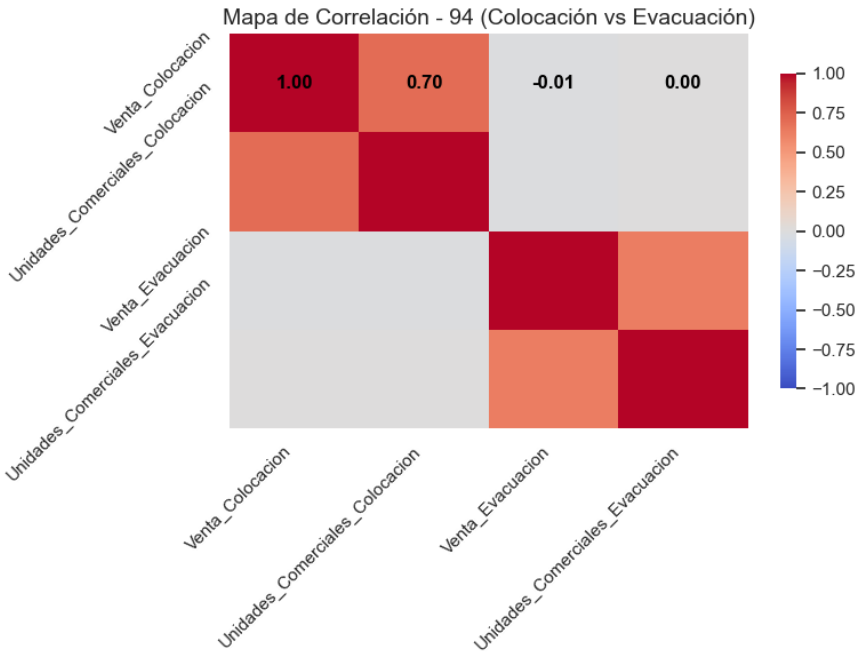
	Venta_Evacuación_94	Unidades_Evacuación_94	Venta_Evacuación_93	Unidades_Evacuación_93
count	1.003348e+06	1.003348e+06	703.000000	703.000000
mean	1.072136e+05	1.091702e+01	17120.519264	2.364232
std	3.197044e+05	2.994778e+01	42197.363877	7.127999
min	-1.740582e+05	-4.000000e+00	0.000000	0.000000
25%	9.075527e+03	1.000000e+00	0.000000	0.000000
50%	3.380213e+04	3.000000e+00	0.000000	0.060000
75%	1.014998e+05	9.538460e+00	16367.699210	2.000000
max	3.535907e+07	2.715000e+03	407740.733150	102.000000

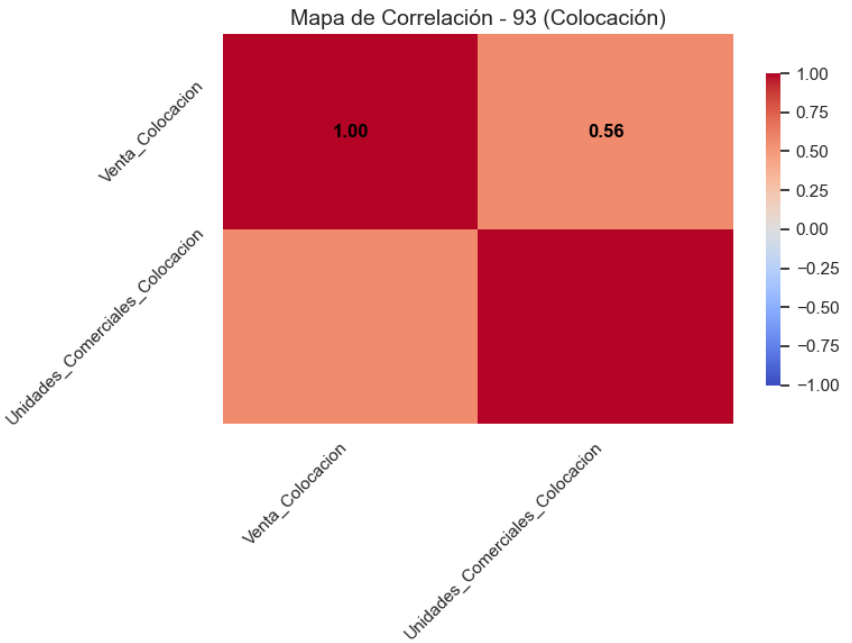
Al realizar el análisis de datos atípicos se encontró: (i) que se contaba con varios datos atípicos o por lo menos por fuera de la media establecida, (ii) en la revisión de intervalos máximos y mínimos existían valores negativos.

Para el primer caso, se decidió continuar trabajando con los datos “Atípicos” porque al comparar la información de Colocación con la información reportada por los negocios en los foros, no se encontró diferencias apreciables por lo que eliminar los atípicos podría ser riesgoso para perder información importante de la venta.

Para el segundo caso se concluyó que los datos negativos se debían a las devoluciones de dinero que existen también en las transacciones con los clientes, lo que también implica un comportamiento normal de la información transaccional de las ventas

4. **Correlación:** De manera inicial las únicas variables numéricas son Venta y Unidades Comerciales.





Para eso, se comparó la correlación existente entre estas variables y lo que se pudo concluir es que:

Primero, al analizar el subcanal 94 se observa que existe una fuerte correlación en la información de venta y unidades comerciales tanto en la información de colocación como de evacuación, sin embargo se observa que no existe correlación entre la información de colocación y evacuación, lo que es extraño y se volvió un punto de atención para la exploración de la información.

Segundo, al analizar el subcanal 93 no se tuvo la necesidad de comparar con evacuación porque esta información de colocación es en esencia la venta que le llega al punto de venta directamente. En este caso se analizar las únicas variables numéricas que tiene y se observa que tienen una correlación que hace que tenga sentido ya que son proporcionales las ventas a las unidades vendidas, tal cual como pasa con el canal 94.

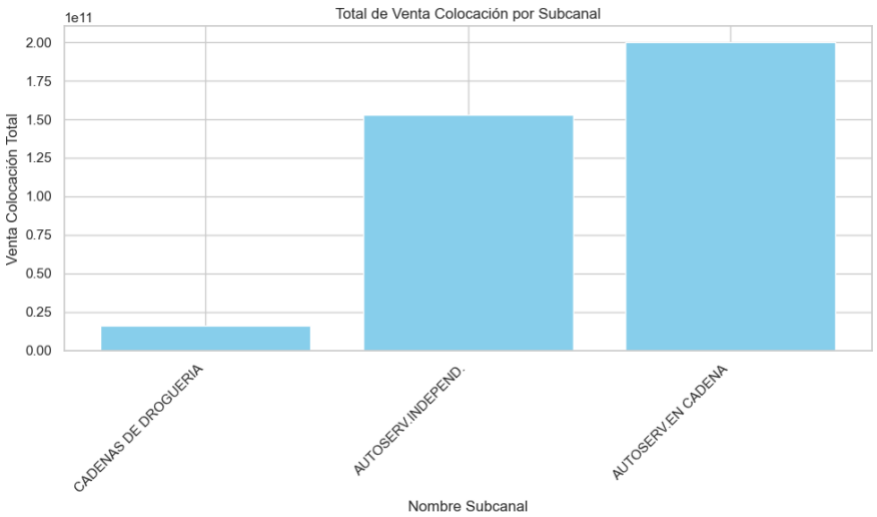
5. **Ajustes de tipos de datos:** Se realizó el ajuste de los tipos de datos de la información para poder tener info categórica como es que es categórica, información numérica como numérica y en este caso el formato fecha como debía estar.

   En este caso se tuvo en cuenta que la información debía quedar agrupada en términos de trimestre (a partir de la información de Fecha)

6. **Creación de variables:** Para poder aportar y enriquecer la información suministrada se realizó el cálculo de dos variables: Frecuencia y Rotación

   Esto bajo la lógica que la frecuencia es la cantidad de veces que se hace un pedido al trimestre

7. **Comportamientos de la data:** Se realizó un análisis general de la información con el propósito de entender el comportamiento de la información y de esta manera poder llegar al dataset que serviría para el modelamiento

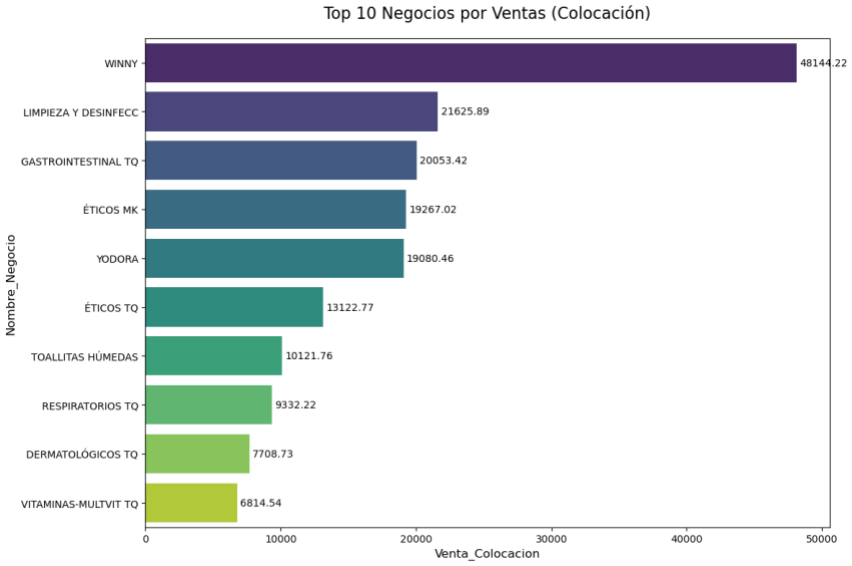


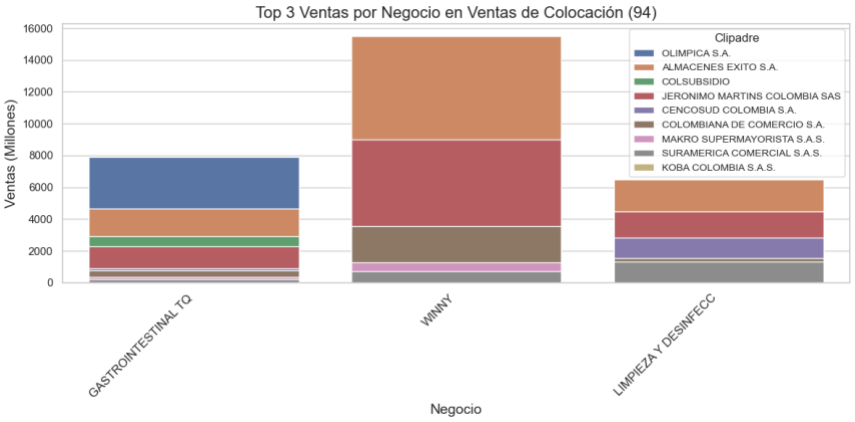
Al ver la información general se puede observar que la venta se encuentra principalmente en el subcanal 94 de grandes cadenas (columna derecha (54%)), seguido de los independientes (41%) y Comfandi (4%) de último.

Posteriormente se realizó un análisis por Subcanal (94 y 93) principalmente y dentro de cada uno se analizó por negocios y por clientes

Para el subcanal 94 entonces analizó la información pero además se realizó la comparación entre la información de colocación y evacuación que era el reto particular de este subcanal

análisis por negocios:

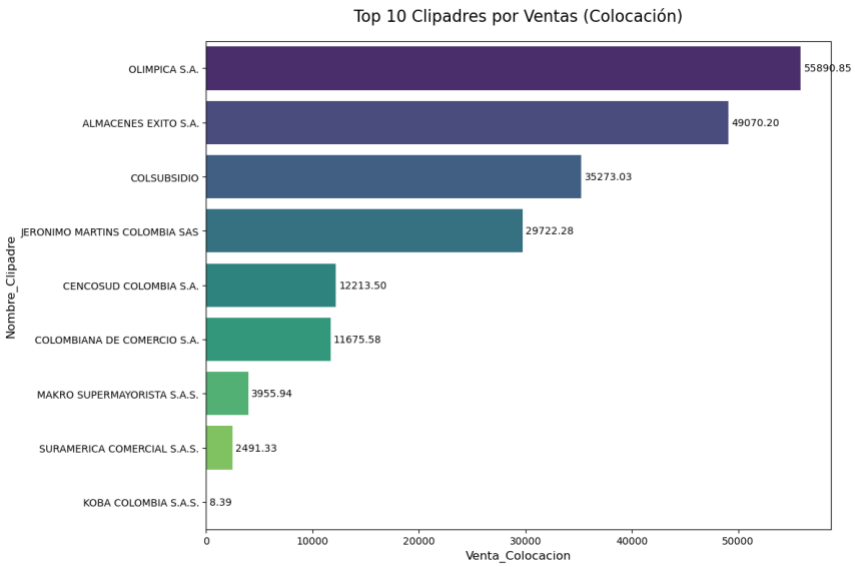


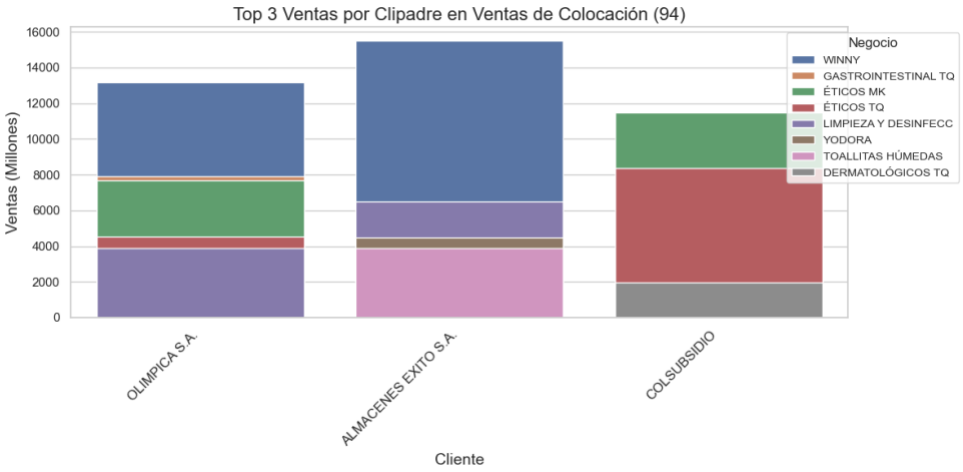


De acá se pudo llegar a algunos hallazgos en cuanto a la información vista por negocios

1. Los negocios de mayor venta en este canal son Winny, Limpieza y Gastrointestinal. Lo cual tiene sentido porque estas grandes cadenas venden principalmente productos de consumo masivo, como lo son los pañales, productos de aseo o limpieza y algunos productos OTC de consumo general.
2. Los clientes más representativos para este negocio son los Almacenes Éxito, Olímpica y Jerónimo Martins.

Análisis por Clientes:

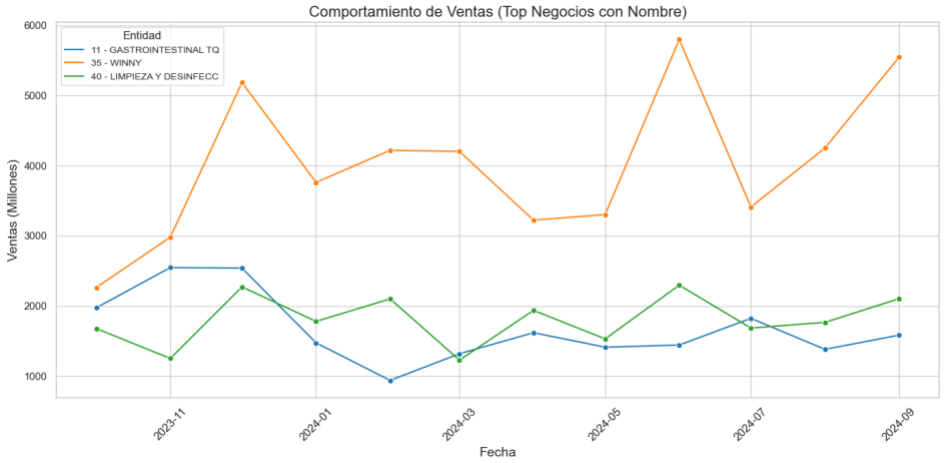


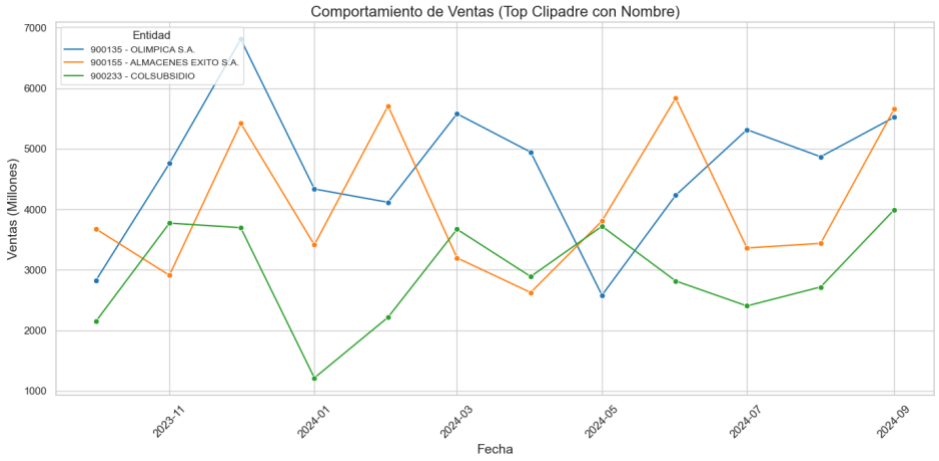


De la mirada puesta desde la información de los clientes, se pudo llegar a otros hallazgos que se correlacionaron la otra mirada también:

1. Los principales clientes de mayor venta en términos generales fueron: Almacenes Éxito, Olímpica, Colsubsidio y Jerónimo Martins
2. Dentro de este foco de clientes se encontró los mismos negocios relevantes: Winny, Limpieza y Gastrointestinal. Sin embargo, se observa que en algunos clientes sobresalen otros solo ahí, como por ejemplo Colsubsidios con el negocio Éticos TQ (que es Farmacéutico) y en Olimpica Éticos MK (que es Farmacéutico) también.

Análisis por Tiempo:





Al revisar la información por tiempo en el subcanal 94 se observó que la información a través del tiempo para las ventas ya sea por negocio o por clientes es estable y consistente en el tiempo, es decir que el comportamiento de las ventas no demuestra algún pico o comportamiento anormal.

Finalmente para el subcanal 94 se realizó una comparación de consistencia en la información de la data de colocación vs la de evacuación, debido a que esta no debería ser tan distinta (al rededor de un 5%)

	Cod_Negocio	Nombre_Negocio	Colocacion_Total_Millones	Evacuacion_Total_Millones	Diferencia_%
0	1	ÉTICOS TQ	13122.772020	3391.579824	74.155005
1	7	ÉTICOS MK	19267.018267	5784.917334	69.975025
2	8	OFTALMOLÓGICOS TQ	1680.934867	325.004866	80.665231
3	10	DERMATOLÓGICOS TQ	7708.731095	3708.389697	51.893643
4	11	GASTROINTESTINAL TQ	20053.420381	11128.220029	44.507122
5	13	VITAMINAS-MULTVIT TQ	6814.535361	3858.089995	43.384401
6	18	RESPIRATORIOS TQ	9332.215719	5153.421297	44.778159
7	19	ALIVIO DOLOR TQ	5663.629258	3272.987265	42.210425
8	21	CUIDADO DE HERIDA TQ	6114.169147	4534.329898	25.838985
9	33	YODORA	19080.464738	12178.339279	36.173781

	Cod_Clipadre	Nombre_Clipadre	Colocacion_Total_Millones	Evacuacion_Total_Millones	Diferencia_%
0	900135	OLIMPICA S.A.	55890.852050	38830.680429	30.524086
1	900155	ALMACENES EXITO S.A.	49070.203288	33423.030117	31.887321
2	900233	COLSUBSIDIO	35273.028264	0.000000	0.000000
3	900489	COLOMBIANA DE COMERCIO S.A.	11675.582900	9593.644814	17.831556
4	900720	MAKRO SUPERMAYORISTA S.A.S.	3955.939559	0.000000	0.000000
5	901652	CENCOSUD COLOMBIA S.A.	12213.503354	9383.883410	23.167963
6	904741	SURAMERICA COMERCIAL S.A.S.	2491.326804	27.839318	98.882551
7	907945	KOBA COLOMBIA S.A.S.	8.391570	0.000000	0.000000
8	908528	JERONIMO MARTINS COLOMBIA SAS	29722.278335	16313.523138	45.113484

En este sentido se evidenció que al hacer la comparación se tenían porcentajes demasiado altos que llegaban hasta el 80% de diferencia por negocio y esto ya empezó a sonar muy extraño en cuanto a la coherencia de la información.

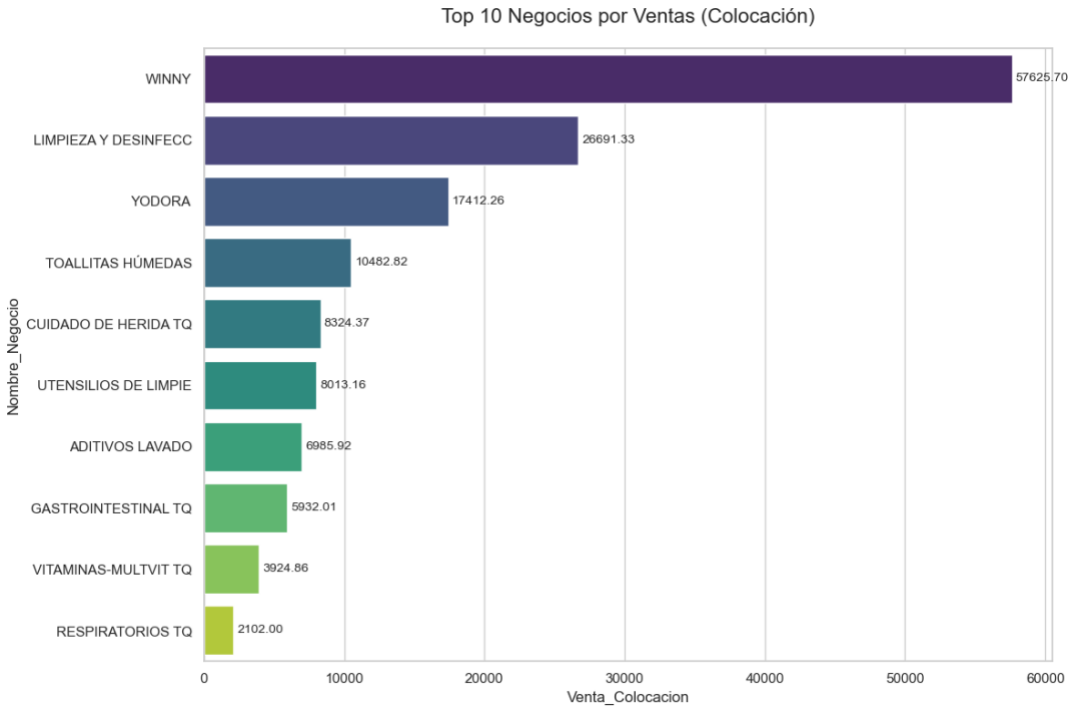
Esto al hacer un doble clic a la información únicamente de evacuación por negocio en el tiempo, dió como resultado la identificación de vacíos o faltantes en la información

Fecha	Cod_Negocio	Nombre_Negocio	2023-10-01 00:00:00	2023-11-01 00:00:00	2023-12-01 00:00:00	2024-01-01 00:00:00	2024-02-01 00:00:00	2024-03-01 00:00:00	2024-04-01 00:00:00	2024-05-01 00:00:00
0	1	ÉTICOS TQ	0.000000e+00	0.000000e+00	4.466156e+08	7.537180e+07	2.306552e+08	2.326760e+08	3.405705e+08	4.736564e+08
1	7	ÉTICOS MK	2.378964e+06	2.407645e+06	8.152200e+08	1.195354e+08	4.564708e+08	3.887216e+08	5.120002e+08	7.858699e+08
2	8	OFTALMOLÓGICOS TQ	0.000000e+00	0.000000e+00	5.085259e+07	5.638215e+06	2.342374e+07	2.418117e+07	3.641469e+07	4.654090e+07
3	10	DERMATOLÓGICOS TQ	1.298254e+08	5.069499e+07	2.993932e+08	2.882344e+08	2.683298e+08	3.179644e+08	2.655020e+08	3.144042e+08
4	11	GASTROINTESTINAL TQ	1.370848e+08	1.021465e+08	1.663151e+09	6.244010e+08	6.303722e+08	8.723208e+08	7.708716e+08	1.157119e+09
5	13	VITAMINAS-MULTVIT TQ	1.919217e+08	5.346650e+07	3.446627e+08	3.952577e+08	3.432818e+08	3.619338e+08	2.604352e+08	3.408978e+08
6	18	RESPIRATORIOS TQ	4.741582e+07	4.707379e+07	5.192738e+08	3.424885e+08	3.201844e+08	4.109046e+08	3.601881e+08	5.498337e+08
7	19	ALIVIO DOLOR TQ	2.485175e+07	2.972499e+07	2.717268e+08	1.776908e+08	2.226679e+08	2.849573e+08	2.910586e+08	3.781312e+08
8	21	CUIDADO DE HERIDA TQ	1.132356e+08	1.197381e+08	4.136516e+08	3.428213e+08	2.853542e+08	4.891114e+08	3.844343e+08	3.970846e+08
9	33	YODORA	6.233894e+08	1.910821e+08	7.610523e+08	1.099724e+09	1.037631e+09	1.011751e+09	7.331222e+08	8.124257e+08
10	35	WINNY	1.133449e+09	4.879243e+08	2.681721e+09	2.365383e+09	2.011524e+09	2.714473e+09	2.214352e+09	2.630851e+09
11	38	CONTENT	0.000000e+00	0.000000e+00	2.853488e+07	2.105915e+07	2.021151e+07	3.092504e+07	3.040742e+07	3.292027e+07
12	40	LIMPIEZA Y DESINFECC	6.061455e+08	2.671503e+08	1.061472e+09	1.240608e+09	1.112282e+09	1.323809e+09	1.084748e+09	1.115727e+09
13	41	ADITIVOS LAVADO	2.301337e+08	8.595925e+07	3.244641e+08	4.718748e+08	3.486873e+08	4.515166e+08	3.125617e+08	2.974444e+08
14	45	TOALLITAS HÚMEDAS	3.113462e+08	1.175584e+08	5.108310e+08	6.488555e+08	6.066714e+08	5.833769e+08	4.809974e+08	4.064679e+08

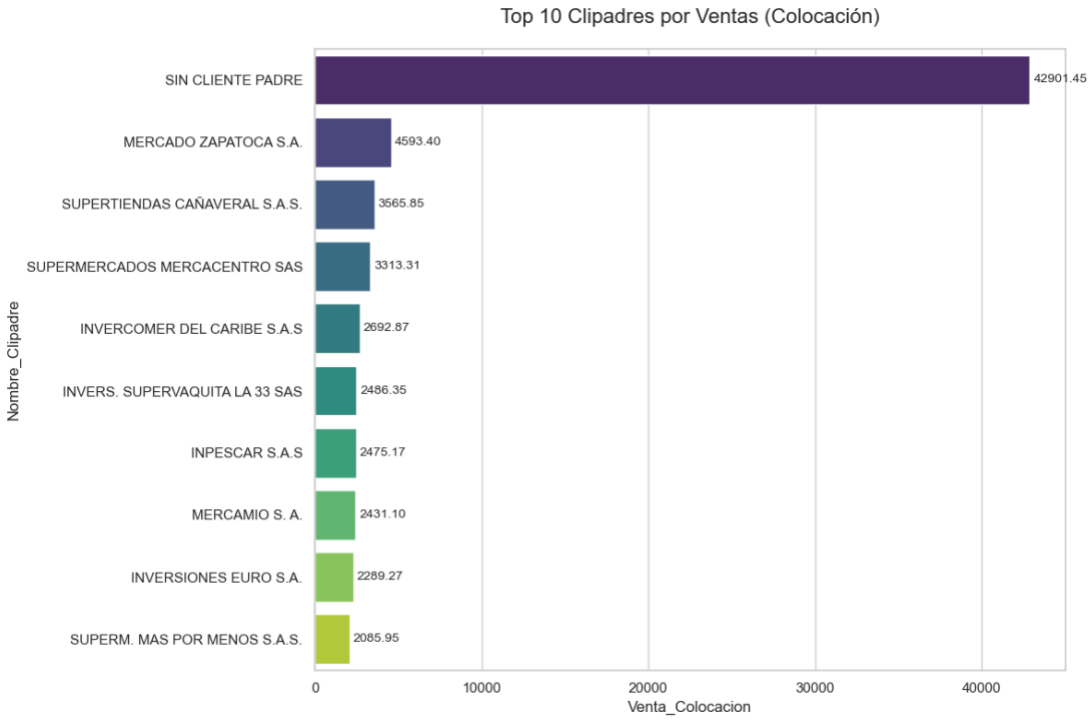
En ese sentido se revisó con el área encargada y efectivamente el sistema tiene errores y la información de evacuación no está completa y por lo tanto no es posible utilizarla para el ejercicio

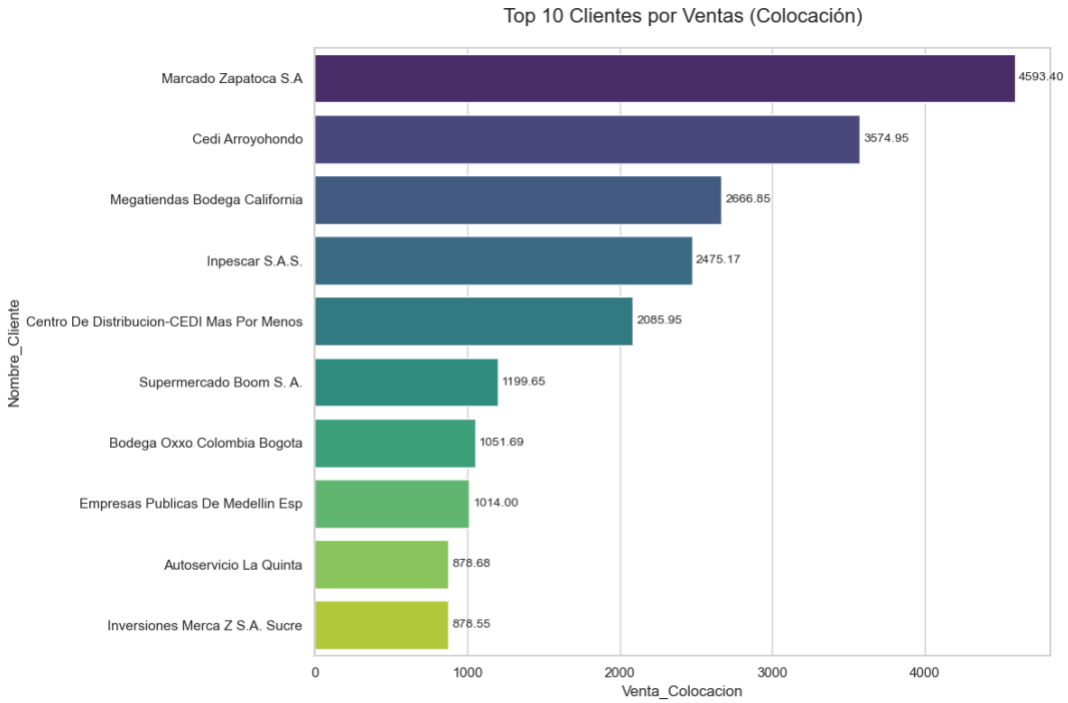
Por otro lado para el canal 93, se realizó un análisis similar pero con la particularidad que en este caso no era necesaria hacer una comparación con la información de evacuación y adicional en este caso no contábamos con agrupaciones de clientes ya que la mayoría de sus clientes son directamente el punto de venta donde se vende

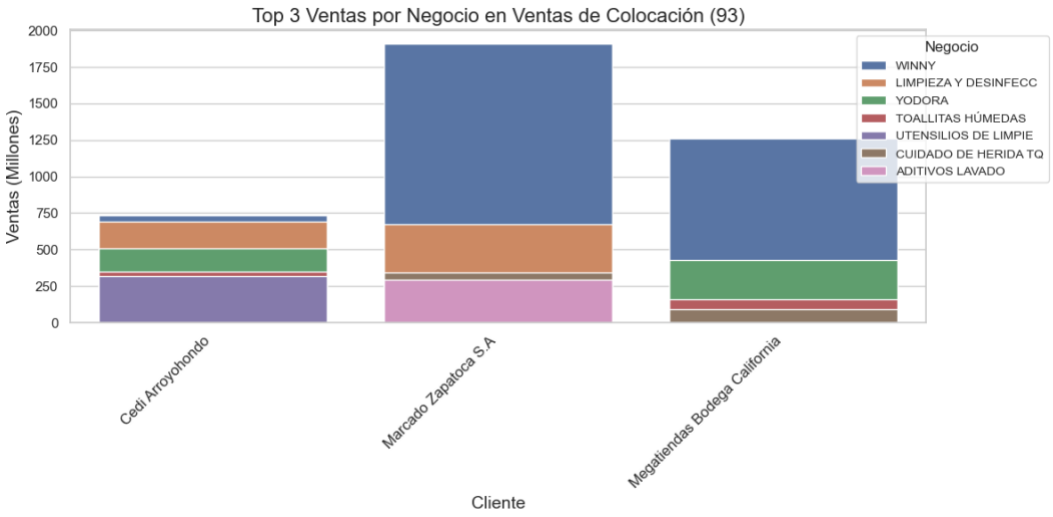
Análisis por Negocio:



En este análisis se pudo ver que Winny y Limpieza también son los negocios más representativos en este subcanal, pero ya en tercero se encuentra Yodora. En este caso los 3 son negocios de consumo masivo y esto tiene sentido que sean los negocios principales en estos clientes que principalmente le apuntan a este nicho de mercado.







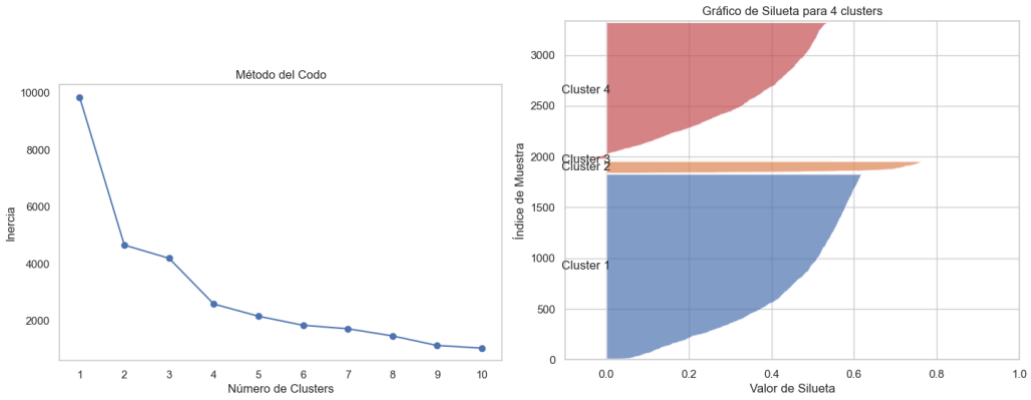
Al analizarlo en función de los clientes, se evidencia que efectivamente en este subcanal los clientes en su mayoría no tienen cliente agrupador/padre y esto es porque lo que se vende al cliente es directamente al punto de venta. Es por esto que es posible asumir de manera correcta que con la venta de colocación es suficiente aproximación a la venta en el punto de venta

Adicionalmente, se analizó la información en detalle de los 3 clientes más representativos y se observa en concordancia con el análisis por negocio que Winny es el que tiene mayor importancia en ventas, luego el negocio de limpieza y luego el de yodora.

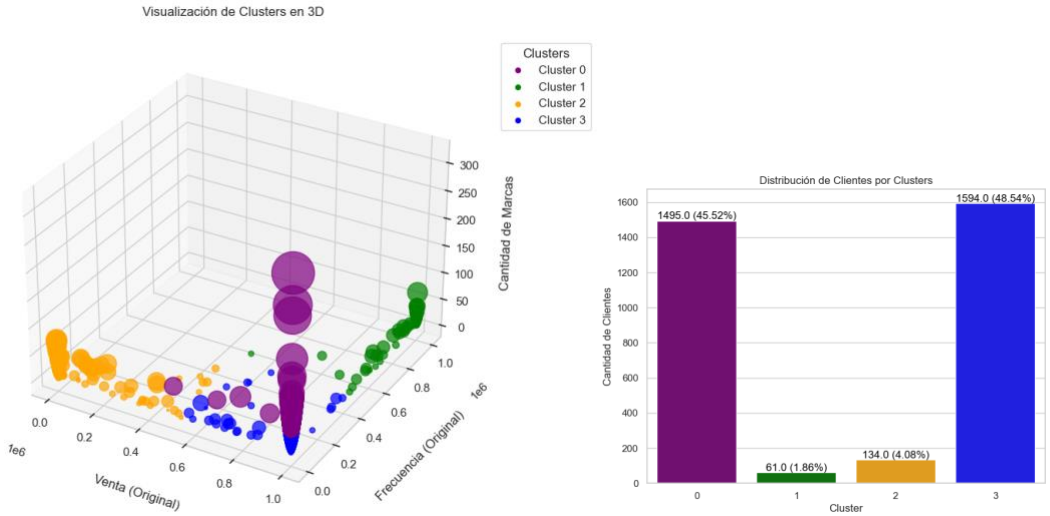
8. clustering para exploración: Debido a que el subcanal 93 no tiene limitantes de información, se optó por iniciar el proceso de modelamiento con este subcanal a través de un análisis inicial por clustering donde se construyó el primer dataset para hacer una segmentación inicial de la siguiente manera:

```
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Cod_Cliente      3284 non-null   int64
 1   Nombre_Cliente   3284 non-null   object
 2   Venta            3284 non-null   float64
 3   Frecuencia       3284 non-null   float64
 4   Cantidad_Marcas  3284 non-null   int64
dtypes: float64(2), int64(2), object(1)
```

De esta manera se inició identificando el número de clusters óptimos para esta primera aproximación a través de gráfica del codo y silhouette y en resumen se decidió modelar con 4 clusters por KNN



Como resultado se pudo observar la clusterización de acuerdo a estas 3 variables de interés: Venta, Frecuencia de compra y Cantidad de marcas vendidas y la composición de clientes resultante en cada uno de los clusters



Al analizar cada clusters descriptivamente frente a sus variables relevantes se obtuvo lo siguiente:

Venta

	Estadistica	0	1	2	3
0	count	1.495000e+03	6.100000e+01	1.340000e+02	1.594000e+03
1	mean	8.214995e+07	2.715223e+07	-6.151082e+04	1.774212e+07
2	std	2.111013e+08	1.348609e+08	5.750281e+05	2.884020e+07
3	min	5.084383e+05	5.464374e+05	-3.829758e+06	5.740581e+05
4	25%	2.034694e+07	2.218401e+06	-2.882991e+04	5.745474e+06
5	50%	3.993128e+07	3.743955e+06	5.406425e+03	1.118139e+07
6	75%	8.111242e+07	9.561014e+06	1.490308e+05	2.024777e+07
7	max	4.593400e+09	1.051692e+09	5.401840e+05	7.700128e+08

Frecuencia

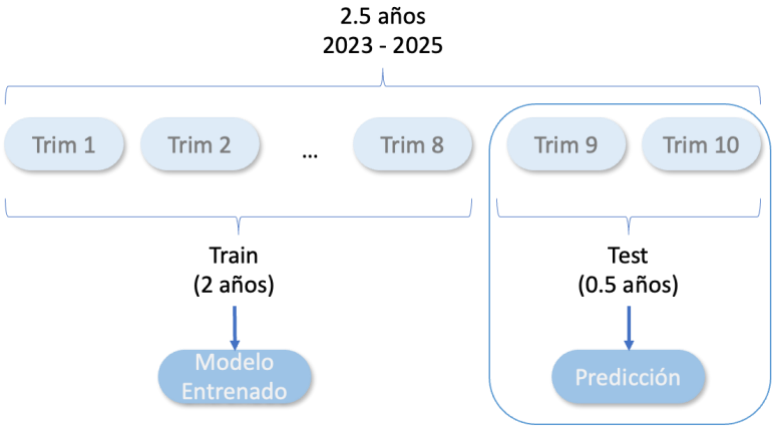
	Estadistica	0	1	2	3
0	count	1495.000000	61.000000	134.000000	1594.000000
1	mean	0.884966	0.959300	0.237196	0.860644
2	std	0.085746	0.059799	0.368480	0.100104
3	min	0.000000	0.650000	0.000000	0.000000
4	25%	0.854928	0.950000	0.000000	0.821403
5	50%	0.902834	0.977143	0.000000	0.886955
6	75%	0.938144	1.000000	0.471154	0.924971
7	max	0.992024	1.000000	1.000000	1.000000

Cantidad de marcas

	Estadistica	0	1	2	3
0	count	1495.000000	61.000000	134.000000	1594.000000
1	mean	59.420736	20.836066	22.917910	32.550816
2	std	14.759934	11.498087	21.285403	9.218843
3	min	46.000000	5.000000	1.000000	2.000000
4	25%	51.000000	14.000000	6.000000	27.000000
5	50%	57.000000	18.000000	14.000000	34.000000
6	75%	65.000000	29.000000	38.750000	40.000000
7	max	317.000000	69.000000	74.000000	45.000000

3. Modelamiento

Cómo diseño de modelos a abordar en el proyecto se realizó la investigación pertinente y se definió explorar 2 tipos de métodos de aprendizaje no supervisado: Modelo híbrido LightFM y Modelo FP Growth



En general el proyecto realizó una implementación de dos modelos donde la información obtenida por el sistema transaccional de ventas se partió en dos: 2 años de información como train y 2 semestres actuales como base de validación para retar los resultados obtenidos por los distintos modelos

Validación de los modelos

los modelos se retaron a través de 2 métricas principalmente utilizadas en este tipo de contexto y variables de venta y unidades comerciales

Hit rate: Mide la proporción de clientes para los cuales al menos uno de los productos recomendados fue efectivamente comprado.

$$\text{Hit Rate} = \frac{\text{Número de clientes con al menos un acierto}}{\text{Número total de clientes evaluados}}$$

Y en la literatura se especifican los valores esperados o deseables para la evaluación de los modelos de recomendación.

Valor	Interpretación
>=0.30	Excelente
0.15 – 0-30	Aceptable
< 0.15	Débil

En algunos casos donde se utilizó está métrica a nivel de cliente y la recomendación realizada para este cliente en particular se utilizó una medición ligeramente distinta, donde lo que se buscaba es identificar el porcentaje de aciertos sobre el total de productos recomendados, esto con el propósito de cuantificar que tan preciso fue el modelo para recomendar productos que efectivamente el cliente terminó comprando en realidad

$$\text{Precision} = \frac{\text{Número de recomendaciones acertadas}}{\text{Número total de recomendaciones realizadas}}$$

Es decir se usó un hit rate como nivel de precisión en ese nivel de detalle y se utilizó la misma escala de interpretación de su resultado porque es mucho más ácida que la escala que la literatura propone para modelos de recomendación, donde recomiendan que un buen modelo de recomendación tiene una precisión del 30%, sin embargo debido a la naturaleza de este negocio, donde no acertar en una marca implica millones de pérdidas, consideré importante mantener una escala más ácida que forzara evaluar las recomendaciones de forma más cercana a un 100%

Sales Coverage: Es la proporción de las ventas totales que están cubiertas por los productos recomendados. Es decir, que mide en cuánto volumen de ventas se cubren los productos que el modelo recomienda

$$\text{Sales Coverage} = \frac{\text{ventas de productos recomendados}}{\text{ventas totales del cliente}}$$

Al igual que con el hit rate, la literatura específica los valores esperados o deseables para la evaluación de los modelos

Valor	Interpretación
>=0.20	Alta Cobertura
0.10 – 0.20	Media Cobertura
< 0.10	Baja Cobertura

Adicional a lo anterior se utilizaron las métricas específicas de cada modelo para poder establecer que tan buena o no era la recomendación de producto. Estas métricas se explicitan mas adelante en la fase de resultados de cada modelo para poder analizar e interpretar lo que se obtuvo.

El hit rate y el Sales Coverage se dejan como métricas de medición propuestas a futuro para evaluar la efectividad de la implementación de las recomendaciones que el modelo haga y en las corridas del modelo sirven para ver si el modelo está recomendando productos nuevos.

3.1 LightFM (Modelo Híbrido para Sistemas de Recomendación)

Primero a modo de contexto, el algoritmo **LightFM** es una técnica híbrida de recomendación que combina enfoques colaborativos y de contenido. Esto significa que se fundamenta en la factorización de matrices, donde se realiza la asignación de algo denominado “representaciones latentes” que son vectores (es decir representaciones numéricas en el espacio, donde "latente" significa que no son características explícitas (por ejemplo variables como edad, precio, color…), sino que el algoritmo las aprende automáticamente a partir de los datos de interacción (quién compró qué, quién hizo clic en qué) - es decir que el modelo asocia características) tanto para los clientes como para las marcas de los productos.

El modelo calcula la **similitud** o **afinidad** entre un cliente y una marca de producto haciendo un **producto punto** entre estos vectores, como se ve en la siguiente fórmula general:

$$\hat{r}_{ui} = U[u] \cdot V[i] + b_u + b_i$$

U[u] = Perfil del cliente, construido a partir de su historial de compras y variables como rotación, ventas, etc.
V[i] = Perfil de la marca de producto, basado en atributos implícitos como su rotación, unidades vendidas, ventas totales, etc.

donde, para cada cliente $U[u]$ y cada marca de producto i $V[i]$ se calcula de alguna manera la predicción de "afinidad" entre un cliente y una marca de producto a través de este producto punto, teniendo en cuenta sus propios sesgos para que matemáticamente la predicción considere un margen de error.

El algoritmo utilizado en este modelo cuenta con distintas funciones de pérdida diseñadas específicamente para escenarios con **datos implícitos**, es decir, situaciones en las que no se dispone de valoraciones explícitas de los usuarios, sino únicamente de registros de interacción (como compras, clics o visualizaciones). Entre las funciones disponibles se destacan **BPR (Bayesian Personalized Ranking)** y **WARP (Weighted Approximate-Rank Pairwise loss)**.

Esto lo hace de la siguiente manera: primero se calcula la "afinidad" mencionada en la fórmula a cada marca frente a cada cliente (el objetivo es que sea mayor), luego se hace un sistema de ranking donde de a pares de afinidades se verifica cual de las dos afinidades tiene el mejor puntaje y el menor se va penalizando en función de la cantidad de ítems que se necesitaron para encontrar la mayor afinidad.

Para poder implementar este modelo se hizo necesario realizar los siguientes pasos:

Paso 1: Preparación de los datos

Todo inició con la construcción del DataFrame que contiene, por cliente, por producto y por mes: venta, unidades comerciales y rotación trimestral

Como el modelo LightFM trabaja con "interacciones" entre la marca y los clientes donde un cliente compra un producto, entonces se hizo necesario representarlo en una matriz dispersa donde las filas eran los clientes y las columnas las marcas de los productos.

Paso 2: Realizar PCA para poder estandarizar la data

Estas variables se estandarizaron con StandarScaler y se redujo su dimensionalidad usando **Análisis de Componentes Principales (PCA)**. Como resultado, lo que se obtuvo fueron dos nuevas variables, PCA_1 y PCA_2, que resumían el comportamiento del producto en dos dimensiones numéricas. Esto fue realmente útil

sobre todo para poder estandarizar las escalas numéricas de las variables que no necesariamente tenían dimensiones similares

<u>Paso 3:</u> Realizar modelo LightFM para encontrar interacciones relevantes entre las marcas y los clientes

Se entrenó el modelo para que estas puntuaciones fueran altas para interacciones reales y bajas para interacciones negativas o inexistentes, según la función de pérdida elegida (warp en este caso).

- Si los vectores de cliente y producto apuntan en **la misma dirección**, el producto punto es **grande y positivo** ⇒ **alta afinidad**
- Si están en **direcciones opuestas**, el valor es negativo ⇒ **baja afinidad**
- Si son ortogonales (no relacionados), el valor será cercano a 0 ⇒ **afinidad neutra o irrelevante**

Resultados - subcanal 94

238 Clientes – 10 recomendaciones principales

Hit Rate	Sales Coverage
26.05%	0.18%

Validación del Modelo

5 Clientes Top de Ventas

Clipadre	Cliente	
Olímpica	Olímpica Bodega 8024	❌
Éxito	Cedi Vegas # 020	✅
Cencosud	Plataforma Cencosud Bogotá	❌
Colombiana de Comercio	Alkosto Avenida 68	❌
Jerónimo Martins	Jerónimo Martins Cedi Montería	✅

Hit Rate	Sales Coverage
40%	0%

Validación del Modelo

5 Clientes Top de Ventas – 10 recomendaciones principales

Clipadre	Cliente	Hit Rate	Sales Coverage
Olímpica	Olímpica Bodega 8024	0%	0%
Éxito	Cedi Vegas # 020	10%	0%
Cencosud	Plataforma Cencosud Bogotá	0%	0%
Colombiana de Comercio	Alkosto Avenida 68	0%	0%
Jerónimo Martins	Jerónimo Martins Cedi Montería	20%	0%

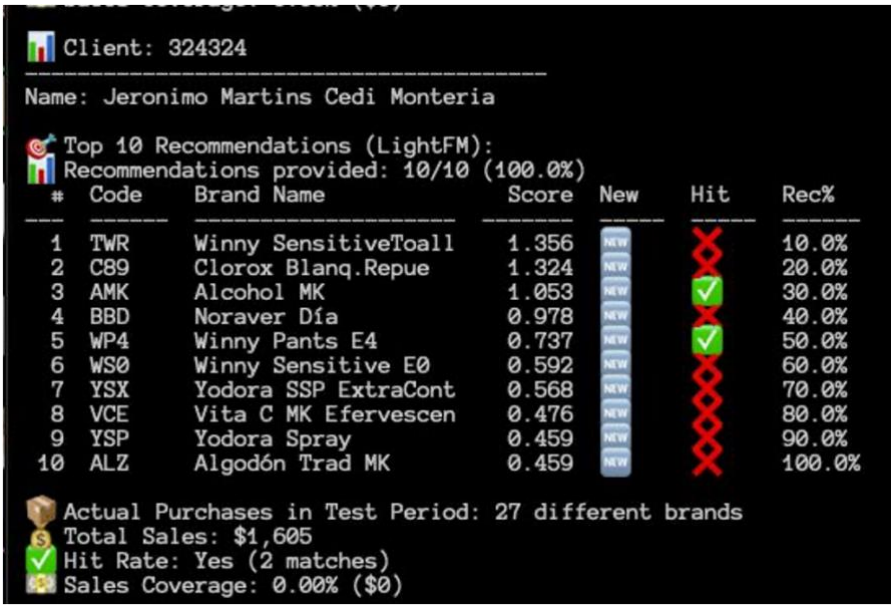
Validación del Modelo

Predicción

Clipadre	Cliente	Rec. 1	Rec. 2	Rec. 3	Rec. 4	Rec. 5	Score
Olímpica	Olímpica Bodega 8024	C78-Limpiadores Bombril	WS1-Winny sensitive E1	NTD-NITAZOXAMK100/5MLSUS	C89-CLOROX BLANQ. REPUE	AXP-ALTEX PREVIENE	-
Éxito	Cedi Vegas # 020	VCG - Vitamina C MK Gotas	C78 - Limpiadores Bombril	CCS - Clotrimaz.Corp.Soluc	BK2 - Blankísima 2%	YOD - Yodopovidona MK	-
Cencosud	Plataforma Cencosud Bogota	CPC - Crema N4 Protect	CPC - Crema N4 Protect	CLG - Clotrimazol MK 2%	KOL - Kola Granulada MK	AX2 - Amoxic.MK 125MG/5ML	-
Colombiana de Comercio	Alkosto Avenida 68	ESL - Esponjillón	INX - Inoxidables	C78 - Limpiadores Bombril	WG6 - Winny Ultra Gold E6	WS2 - Winny Sensitive E2	-
Jerónimo Martins	Jeronimo Martins Cedi Monteria	TWR - Winny SensitiveToal	C89 - Clorox Blanq.Repuest	AMK - Alcohol MK	BBD - Noraver Día	WP4 - Winny Pants E4	+

Para este subcanal el modelo terminó recomendando productos para 1/5 clientes evaluados a partir de la siguiente escala de métrica Score del modelo:

Score	Interpretación
>=1.5	Alta afinidad
0.5 – 1.5	Afinidad moderada
0 – 0.5	Afinidad Baja
<0	Rechazo*



En este caso se obtuvo varias marcas de productos con una afinidad moderada y casi alta que implicaría una apuesta para un piloto que pueda luego corroborarse con con las métricas de hit rate y sales coverage para ver que tan acertada fue la recomendación.

Resultados - subcanal 93

3.003 Clientes – 5 recomendaciones principales

Hit Rate	Sales Coverage

Validación del Modelo

5 Clientes Top de Ventas

Cliente	
Mercados Lemar	❌
Servialgusto Condominio Mediterrane	❌
Euro jardines de llano grande	❌
Supermercado Popular 01	❌
Supermercado Mercacentro #22	✅

Hit Rate	Sales Coverage
20%	0.12%

Validación del Modelo

5 Clientes Top de Ventas – 5 recomendaciones principales

Cliente
Mercados Lemar
Servialgusto Condominio Mediterrane
Euro jardines de llano grande
Supermercado Popular 01
Supermercado Mercacentro #22

Hit Rate	Sales Coverage
0%	0%
0%	0%
0%	0%
0%	0%
20%	1.5%

Validación del Modelo

Predicción

Cliente	Rec. 1	Rec. 2	Rec. 3	Rec. 4	Rec. 5	Score
Mercados Lemar	C85 - Clorox Mancha Garra	YDR - Yodora Derma Rll	YTR - Yodora Total Rll	ABS - Absorbente	C73 - Limpido Mancha Garra	+/-
Servialgusto Condominio Mediterrane	PC7 - Pañal Content Ultras	C89 - Clorox Blanq.Repue	CPB - Venditas Standard	BBD - Noraver Día	C76 - Limpido Blanq.Repue	+/-
Euro jardines de llano grande	421 - Pinesol	801 - Alambre	IBF - Ibuflash Forte	CMU - Crema N4 Multiusos	YOD - Yodopovidona MK	-
Supermercado Popular 01	WS2 - Winny Sensitive E2	05E - Yodosalil	BLU - Bonfiest	CAP - Cinta Papelería	421 - Pinesol	-
Supermercado Mercacentro #22	BBD - Noraver Día	ESP - Esponjilla	ABS - Absorbente	WS1 - Winny Sensitive E1	PC7 - Pañal Content Ultras	-

Para el subcanal 93, se observa una recomendación de marcas de productos para 2/5 de clientes. Estos basados en una recomendación con Score positivo

```
Client: 256684
----------------------------------------
Name: Mercados Lemar

Top 5 Recommendations (LightFM):
Recommendations provided: 5/5 (100.0%)
  #  Code    Brand Name              Score  New  Hit   Rec%
---  ------  --------------------  -------  ---  ---  ------
  1  C85     Clorox Mancha Garra     0.321  NEW  ✗    20.0%
  2  YDR     Yodora Derma Rll        0.16   NEW  ✗    40.0%
  3  YTR     Yodora Total Rll        0.085  NEW  ✗    60.0%
  4  ABS     Absorbente             -0.063  NEW  ✗    80.0%
  5  C73     Limpido Mancha Garra   -0.095  NEW  ✗    100.0%

Actual Purchases in Test Period: 49 different brands
Total Sales: $15,340
Hit Rate: No (0 matches)
Sales Coverage: 0.00% ($0)
```

Tomando de ejemplo Mercados Lemar, el Score de afinidad es más de una afinidad baja y pues ahí tocaría revisar si vale la pena o no iniciar estrategias comerciales que incluyan estos productos.

Con estos resultados se pudo evidenciar un potencial latente de este modelo pues al utilizar el modelo con el recurso computacional limitado, demostró ser altamente eficiente, pues pudo correrse en su totalidad en un tiempo realmente corto.

3.2 FP-Growth (Reglas de Asociación)

Este modelo está diseñado para descubrir asociaciones frecuentes entre ítems en un conjunto de datos transaccionales. Se basa en el análisis de co-ocurrencias, identificando qué productos suelen comprarse juntos. En ese sentido, la propuesta

va encaminada a realizar estas asociaciones dentro de conjunto de productos para cada cliente y encontrar sus patrones de compra

Este modelo no utiliza una única fórmula matemática explícita como el modelo Light FM que se basa en funciones objetivo como por ejemplo warp, sino que se basa en una estrategia computacional para encontrar patrones frecuentes en transacciones sin generar combinaciones innecesarias.

En ese sentido, se apoya en métricas fundamentales durante el proceso de minería de reglas de asociación:

1. Soporte (Support)

Indica la frecuencia con la que ocurre un conjunto de ítems (productos) en todas las transacciones.

$$\text{Support}(X) = \frac{\text{Número de transacciones que contienen } X}{\text{Número total de transacciones}}$$

Ejemplo:

Si 20 clientes de un total de 100 compraron tanto la Marca A como la Marca B, entonces:

$$\text{support}(A, B) = \frac{20}{100} = 0.2$$

2. Confianza (Confidence)

Mide la probabilidad de que un cliente compre la marca B dado que ya compró la marca A.

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Ejemplo:

Si 20 clientes compraron A y B juntos, y 25 clientes compraron A, entonces:

$$\text{confidence}(A \Rightarrow B) = \frac{20}{25} = 0.8$$

3. Lift (Elevación)

Esta mide cuánto más probable es que ocurra la compra conjunta de A y B comparado con que ocurran de forma independiente.

$$\text{Lift}(A \Rightarrow B) = \frac{\text{Confidence}(A \Rightarrow B)}{\text{Support}(B)}$$

Ejemplo:

Si:

- confidence(A ⇒ B) = 0.8
- support(B) = 0.4

Entonces:

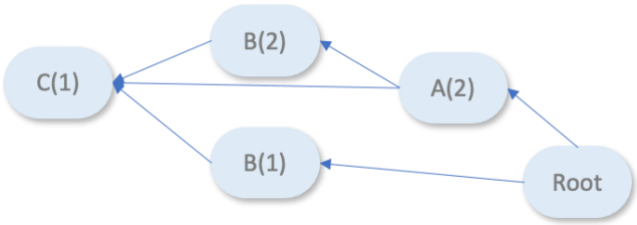
$$\text{lift}(A \Rightarrow B) = \frac{0.8}{0.4} = 2.0$$

Interpretación:

- Lift = 1: A y B son independientes (no hay relación).
- Lift > 1: A y B tienen **relación positiva** (comprar A **incrementa la probabilidad** de comprar B).
- Lift < 1: A y B tienen **relación negativa** (comprar A **reduce la probabilidad** de comprar B).

Este modelo se fundamenta en los siguientes pasos importantes:

1. Preparación de la data en formato de transacciones (cada transacción es la lista de marcas que un cliente compró)
2. Cálculo de la frecuencia de cada producto (**Soporte**)
3. Construcción de un FP-Tree, este construye una estructura en forma de árbol para almacenar las transacciones



- Las transacciones se ordenan por los ítems más frecuentes.
- Se insertan en el árbol compartiendo nodos comunes.
- Cada nodo representa un ítem, y se guarda el conteo de cuántas veces aparece en esa posición del árbol.

4. Extracción de itemsets frecuentes (combinaciones de productos)
5. Generación de reglas de asociación
    - **Confianza** (A → B) = Soporte A U B / Soporte A
    - **Lift** () = Confianza (A → B) / Soporte B

6. Aplicar reglas de asociación en recomendaciones de marcas

Resultados - subcanal 94

238 Clientes – 5 recomendaciones principales

Hit Rate	Sales Coverage
25.21%	0.19%

Validación del Modelo

5 Clientes Top de Ventas

Clipadre	Cliente	
Olímpica	Olímpica Bodega 8024	✗
Éxito	Cedi Vegas # 020	✓
Cencosud	Plataforma Cencosud Bogotá	✗
Colombiana de Comercio	Alkosto Avenida 68	✗
Jerónimo Martins	Jerónimo Martins Cedi Montería	✓

Hit Rate	Sales Coverage
40%	0.07%

Validación del Modelo

5 Clientes Top de Ventas – 5 recomendaciones principales

Clipadre	Cliente	Hit Rate	Sales Coverage
Olímpica	Olímpica Bodega 8024	0%	0%
Éxito	Cedi Vegas # 020	20%	0.07%
Cencosud	Plataforma Cencosud Bogotá	0%	0%
Colombiana de Comercio	Alkosto Avenida 68	0%	0%
Jerónimo Martins	Jerónimo Martins Cedi Montería	20%	0.42%

Validación del Modelo

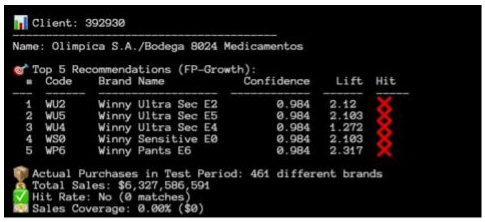
Predicción

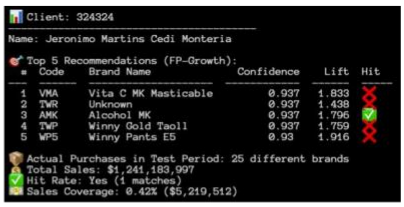
Clipadre	Cliente	Rec. 1	Rec. 2	Rec. 3	Rec. 4	Rec. 5
Olímpica	Olímpica Bodega 8024	YCP - Yodora Cr ClasicaPot	C4O - Crema N4 Original	WU4 - Winny Ultra Sec E4	WU5 - Winny Ultra Sec E5	WU3 - Winny Ultra Sec E3
Éxito	Cedi Vegas # 020	C78 - Limpiadores Bombril	VCG - Vitamina C MK Gotas			
Cencosud	Plataforma Cencosud Bogota	VPV - SIN CATEGORY	CPC - Crema N4 Protect			
Colombiana de Comercio	Alkosto Avenida 68	COB - Copitos Bio MK	COP - Copitos MK	ALZ - Algodón Trad MK	C86 - Clorox Blanq.Botella	YSS - Yodora SSP Suave
Jerónimo Martins	Jeronimo Martins Cedi Monteria	VMA - Vita C MK Masticable	AMK - Alcohol MK	TWP - Winny Gold Taoll	WP5 - Winny Pants E5	YSP - Yodora Spray

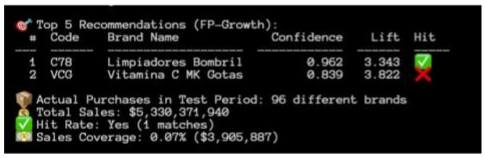
Para interpretar estos resultados se hace necesario entender las métricas propias del modelo, como está a continuación:

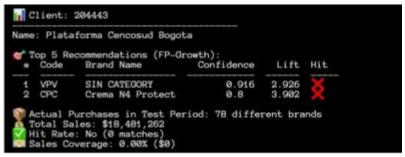
Confianza	Interpretación
>=0.8	Muy Fuerte
0.6 – 0.8	Aceptable
<0.6	Débil

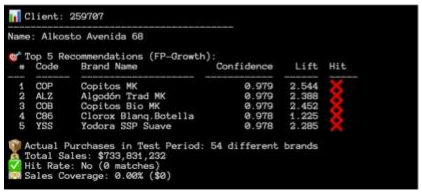
Lift	Interpretación
>2	Muy fuerte asociación
>1	Buena asociación
=1	No hay asociación











Lo que se pudo observar es que este modelo es capaz de encontrar más recomendaciones de marcas de producto y con una confianza muy fuerte, es decir que la gran mayoría aparentan garantizar una apuesta potente de aportar a nuevas estrategias comerciales.

Resultados - subcanal 93

No fue posible obtener resultados para el subcanal 93, porque el recurso computacional no lo permitió. Esto debido a que este modelo es más exigente en cuanto a capacidad de cómputo

Como conclusión se puede decir que este modelo tiene un gran potencial para generar recomendaciones. Sin embargo, utiliza mucho más recurso computacional, y eso no permitió explorar el sub canal 93 debido a que este presenta mucho mayor volumen de datos

Conclusiones

1. El modelo Light FM resultante entregado presenta buenos indicadores Score para generar recomendaciones de productos nuevos que los clientes no han comprado históricamente. Es importante seguir robusteciendo el modelo incorporando más variables no explícitas de las marcas de productos

2.El modelo FPGrowth mostró muy buenos resultados en los indicadores de confianza y lift. Sin embargo, requiere un mayor recurso computacional y por ende no se pudo correr todos los subcanales y todos los clientes. Este modelo fue capaz de generar un mayor número de recomendaciones.

3. El subcanal 91 se quedó por fuera de este ejercicio al igual que la información de evacuación debido a que en el tiempo se realizó el entrenamiento y afinamiento de los modelos con la información principalmente de colocación de los principales subcanales

Limitaciones:

La principal limitante de esta segunda etapa del proyecto fue el poco conocimiento que se tenía sobre los modelos a utilizar que son ***"modelos de asociación"*** de aprendizaje no supervisado, pues el contenido de la maestría se centra principalmente en el aprendizaje supervisado. Por este motivo, la mayor cantidad de tiempo se dedicó a la investigación y entendimiento de los modelos para su uso e implementación con la información del negocio. Adicionalmente, no se contaba con unas métricas preestablecidas y tocó buscar cuales podían adecuarse mejor para rendir cuentas del resultado del modelo/s utilizados durante el proyecto

Trabajo Futuro:

1. Se puede continuar robusteciendo los parámetros del modelo FPGrowth para poder incrementar su eficiencia pues actualmente se tiene un gran consumo de recurso computacional

2. Debido a la naturaleza que tiene el modelo Light FM, es posible incorporar en un siguiente proyecto más variables que no necesariamente son explicitas, como por ejemplo ubicaciones, atributos del producto, etc. Esto permitiría enriquecer la matriz de interacciones e incrementar la capacidad del modelo para captar patrones complejos, fortaleciendo así la calidad de las recomendaciones generadas.

Referencias

1. Ahmad, S., & Pothen, J. (2024). Exploratory Data Analysis and Data Segmentation UsingK-Means Clustering. IEEE Xplore. Ubicado en de https://ieeexplore.ieee.org/document/10183143.

2. Garcia, M., & Lopez, P. (2024). Toward Enhanced Customer Transaction Insights: AnApriori Algorithm-Based Analysis of Sales Patterns. International Journal of AdvancedComputer Science and Applications, 15(2). Recuperado de https://thesai.org.

3. Kula, M. (2015). Metadata Embeddings for User and Item Cold-StartRecommendations. Proceedings of the 2nd Workshop on New Trends in Content-BasedRecommender Systems.

4. Schubert, E., & Rousseeuw, P. J. (2019). Clustering by Fast Search and Find of DensityPeaks. Pattern Recognition, 96, 1-15.

5. Zhou, Y., & Sun, L. (2021). The Role of Sales Channels in the Omnichannel Environment: A Literature Review. Journal of Retailing and Consumer Services, 61.

6. Jannach, D., Adomavicius, G., & Tuzhilin, A. (2016). Recommender Systems – Challenges, Insights and Research Opportunities. *ACM Transactions on Management Information Systems (TMIS)*, 7(1), 1–34. https://doi.org/10.1145/2843948

7. Ricci, F., Rokach, L., & Shapira, B. (2022). *Recommender Systems Handbook* (3rd ed.). Springer. https://doi.org/10.1007/978-1-0716-2197-4
(Capítulos relevantes: Evaluation Metrics for Recommender Systems)

8. Cremonesi, P., Koren, Y., & Turrin, R. (2010). Performance of Recommender Algorithms on Top-N Recommendation Tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (pp. 39–46). ACM. https://doi.org/10.1145/1864708.1864721

9. Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian Personalized Ranking from Implicit Feedback. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*.

10. Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.

11. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 1–38.

Anexos

Notebook del proyecto:
https://drive.google.com/drive/folders/1sbe_JefD3TvH2G8urmpjPogxFz36JKkW?usp=share_link