Informe Final del Proyecto de Data Science

Título del Proyecto: Análisis y Segmentación de Compradores de Productos para Mascotas utilizando Modelos de Clustering

Integrantes del Equipo: Lucas Godoy F, Lucas Goycoolea I, Rodrigo Marquez N

---

1. Introducción

Objetivo Principal: El objetivo de este proyecto es identificar y analizar los diferentes tipos de compradores de productos para mascotas mediante un modelo de clustering basado en un dataset de órdenes históricas. Esto permitirá segmentar a los consumidores según sus patrones de compra y preferencias de productos, proporcionando insights valiosos para optimizar las estrategias comerciales.

Motivación: La correcta segmentación de los clientes es crucial para personalizar las estrategias de marketing y aumentar la eficiencia en la oferta de productos. A través de este proyecto, buscamos proporcionar una metodología robusta y replicable para el análisis de datos de órdenes históricas, contribuyendo a la toma de decisiones informadas en el sector de productos para mascotas.

2. Metodología

2.2 Selección del Dataset

Descripción**:** Se utilizó un dataset que contiene más de 1 millón de filas de órdenes de compra de productos para mascotas. Incluye información detallada sobre los productos comprados, el departamento correspondiente y si el producto es una reorden.

Origen**:** El dataset fue obtenido de Kaggle, y en específico este dataset es de un supermercado no especificado.

2.3 Preparación del Dataset

Depuración de Datos: Del dataset original, eliminamos todos los pedidos que no contenían ningún producto del departamento de mascotas.

Selección de Características: Seleccionamos las características más relevantes para el clustering, incluyendo la categoría de productos y frecuencia de reordenes. Esta selección se basó en su capacidad para capturar patrones significativos.

Análisis Exploratorio de Datos: Utilizamos herramientas como RStudio para realizar un análisis descriptivo inicial, identificando patrones y tendencias clave en los datos. Este análisis incluyó visualizaciones como histogramas, gráficos de barras y diagramas de dispersión para explorar la distribución de órdenes por departamento y la frecuencia de reordenes.

3. Análisis Descriptivo de Datos

3.1 Exploración Inicial

Distribución de Órdenes por Departamento: Analizamos la distribución de órdenes de compra por departamento para identificar cuáles son los más populares. Se observó que los departamentos de "produce", "dairy eggs" y "pets" tienen una alta frecuencia de órdenes, lo que indica una fuerte demanda en estos segmentos. Por otro lado, departamentos como "bulk", "missing", y “other” tienen una menor cantidad de órdenes.

Frecuencia de Reordenes: Investigamos la frecuencia de reordenes de productos, descubriendo patrones de lealtad del cliente. Se encontró que productos del departamento de "produce" tienen una alta frecuencia de reordenes, lo que sugiere que los clientes prefieren productos frescos y tienden a comprarlos repetidamente. Este patrón de lealtad es menos pronunciado en departamentos como "alcohol" y "bulk".

Herramientas Utilizadas: RStudio fue la herramienta principal para el análisis descriptivo. Se escogió para el análisis de datos debido a que RStudio tiene una mayor capacidad para manejar grandes volúmenes de datos y aplicar técnicas de visualización que facilitaron la identificación de patrones y tendencias clave.

4. Desarrollo del Modelo de Clustering

4.1 Selección del Algoritmo

K-means:

- Implementación: Utilizamos el método de codo para determinar el número óptimo de clusters.
- Problema: El gráfico de codo no mostró un punto de inflexión claro, dificultando la determinación del número óptimo de clusters.

Índice de Silueta:

- Implementación: Calculamos el índice de silueta promedio para diferentes números de clusters.
- Problema: La ejecución del cálculo fue demasiado lenta, posiblemente debido al tamaño del dataset.

Bayesian Information Criterion (BIC) usando mclust:

- Implementación: Utilizamos el paquete mclust para ajustar modelos de mezcla de Gaussianas y determinamos el número óptimo de clusters basado en el BIC.
- Problema: El gráfico de BIC sugirió que un solo cluster era el óptimo, lo cual no era útil para el análisis de clusters de diferentes tipos de compradores.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- Implementación: Aplicamos DBSCAN para encontrar clusters basados en la densidad.
- Problema: No hubo suficiente información en los pasos anteriores para evaluar completamente el éxito o fracaso de este método, pero el enfoque parece haber sido eclipsado por problemas previos en la preparación de datos.

4.2 Método utilizado

Proceso de Clustering basado en Distancias:

Método: Finalmente, decidimos agrupar los departamentos manualmente utilizando la información de las distancias con respecto al departamento "pets". Este método se basó en la observación de saltos significativos en las distancias, lo que sugirió agrupaciones naturales en los datos.

1) Recolección y Organización de Datos

Recolección de Datos: Se obtuvo una matriz de distancias que muestra la proximidad entre diferentes departamentos, incluida la distancia con el departamento "pets".

Organización en Excel: Los datos fueron ingresados en una hoja de cálculo de Excel, donde cada celda representa la distancia entre dos departamentos. La columna correspondiente al departamento "pets" se identificó para su análisis.

2) Ordenación de la Columna "Pets"

Seleccionar la Columna: La columna que contiene las distancias entre cada departamento y "pets" fue seleccionada.

Ordenar de Menor a Mayor: Utilizando la función de ordenación de Excel, la columna fue ordenada de menor a mayor para identificar qué departamentos están más cercanos al departamento "pets". Esto facilitó la visualización de las relaciones entre departamentos basadas en sus distancias.

3) Identificación de Saltos Significativos

Análisis de Distancias: Se examinaron las distancias ordenadas para encontrar cambios abruptos, conocidos como "saltos". Estos saltos indican una separación natural entre grupos de departamentos.

Saltos Identificados:

- Entre "dairy eggs" (375) y "beverages" (739): Salto de 364
- Entre "beverages" (739) y "frozen" (1065): Salto de 326
- Entre "frozen" (1065) y "snacks" (1096): Salto de 31
- Entre "snacks" (1096) y "produce" (1402): Salto de 306
- Entre "produce" (1402) y "pantry" (1872): Salto de 470

Interpretación de Saltos: Se observó que los saltos significativos indican dónde los departamentos están más o menos relacionados con "pets", sugiriendo la formación de clusters.

4) Agrupamiento de Departamentos (Clustering)

Definición de Clusters: Basándonos en los saltos identificados, se decidió formar tres clusters:

- Cluster 1: Departamentos muy relacionados con "pets" (distancias pequeñas).
- Cluster 2: Departamentos moderadamente relacionados con "pets" (distancias intermedias).
- Cluster 3: Departamentos menos relacionados con "pets" (distancias grandes).

Asignación de Departamentos a Clusters:

Cluster 1:

- dairy eggs (375)
- beverages (739)
- frozen (1065)
- snacks (1096)

Cluster 2:

- produce (1402)
- pantry (1872)
- household (1913)
- bakery (2292)
- canned goods (2315)
- dry goods pasta (2556)
- deli (2590)
- personal care (2639)
- meat seafood (2740)
- breakfast (2769)

Cluster 3:

- babies (3040)
- international (3083)
- alcohol (3094)
- missing (3245)

- other (3262)
- bulk (3294)

Este agrupamiento se basó en la observación de saltos significativos en las distancias ordenadas, lo que sugiere agrupaciones naturales en los datos.

5. Resultados y Discusión

Clusters Identificados:

- Cluster 1: dairy eggs, beverages, frozen, snacks.
- Cluster 2: produce, pantry, household, bakery, canned goods, dry goods pasta, deli, personal care, meat, seafood, breakfast.
- Cluster 3: babies, international, alcohol, missing, other, bulk.

6. Impacto y Aplicabilidad

Relevancia del Proyecto: Mejora de la segmentación de clientes y optimización de estrategias comerciales.

Potencial de Implementación:

- Estrategias de Marketing Personalizadas: Desarrollar campañas de marketing dirigidas a cada cluster específico para mejorar la efectividad de las promociones y aumentar la satisfacción del cliente.
- Optimización de Inventario: Ajustar el inventario según las preferencias y patrones de compra de cada cluster para minimizar costos y mejorar la disponibilidad de productos populares.

7. Conclusiones

Este proyecto ha demostrado cómo el análisis de datos y la segmentación de clientes pueden proporcionar insights valiosos que mejoran significativamente las estrategias comerciales en el mercado de productos para mascotas. Mediante la implementación de modelos de clustering y la observación de saltos significativos en las distancias entre departamentos, logramos identificar patrones de compra y segmentar a los clientes en grupos distintivos basados en sus comportamientos y preferencias. Esta metodología no solo ha permitido una segmentación precisa, sino que también ha proporcionado un marco adaptable a otros mercados, mostrando la versatilidad y el valor del análisis de clustering para guiar decisiones estratégicas y operativas, desde la personalización de campañas de marketing hasta la optimización de inventario.

8. Anexos

Tabla distancias con departamento pets:

Distancias	pets
pets	0
dairy eggs	375
beverages	739
frozen	1065
snacks	1096
produce	1402
pantry	1872
household	1913
bakery	2292
canned goods	2315
dry goods pasta	2556

deli	2590
personal care	2639
meat seafood	2740
breakfast	2769
babies	3040
international	3083
alcohol	3094
missing	3245
other	3262
bulk	3294