Sistema de Recomendación de Productos para Comercios Locales

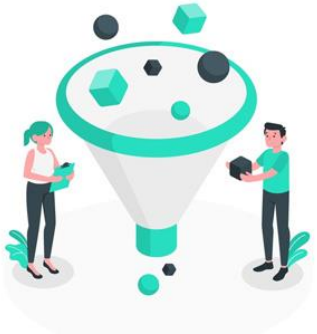
Desafío técnico para optar al cargo Data Scientist

Rindolfo Barra

Septiembre, 2024

Contexto

Mayorista11 busca desarrollar un sistema de recomendación de productos para comercios locales, con objetivo de optimizar ventas y controlar inventario al sugerir productos relevantes.

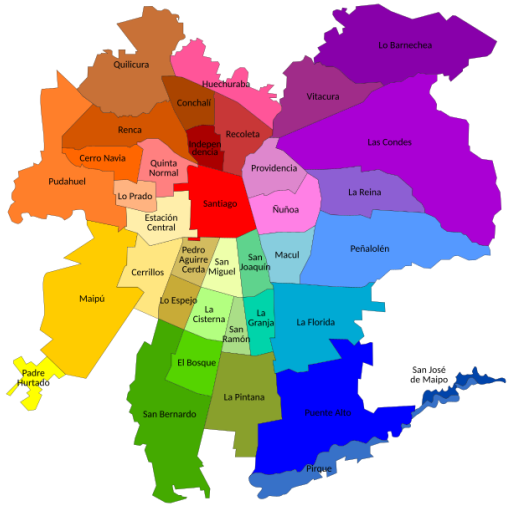


Objetivo

Mejorar la experiencia del cliente y aumentar la retención mediante recomendaciones personalizadas basadas en patrones locales de compra.

Hipótesis Iniciales

1. **Patrones de compra locales**: Priorizar productos que son populares entre comercios similares en la misma comuna.



2. **Historial de compras del comercio**: Tener en cuenta los productos que el comercio ya ha comprado para evitar recomendaciones redundantes.

Descripcion de los datos

- **Comercios**:

  - **id_commerce**: ID único del comercio.
  - **district**: Comuna donde opera el comercio.

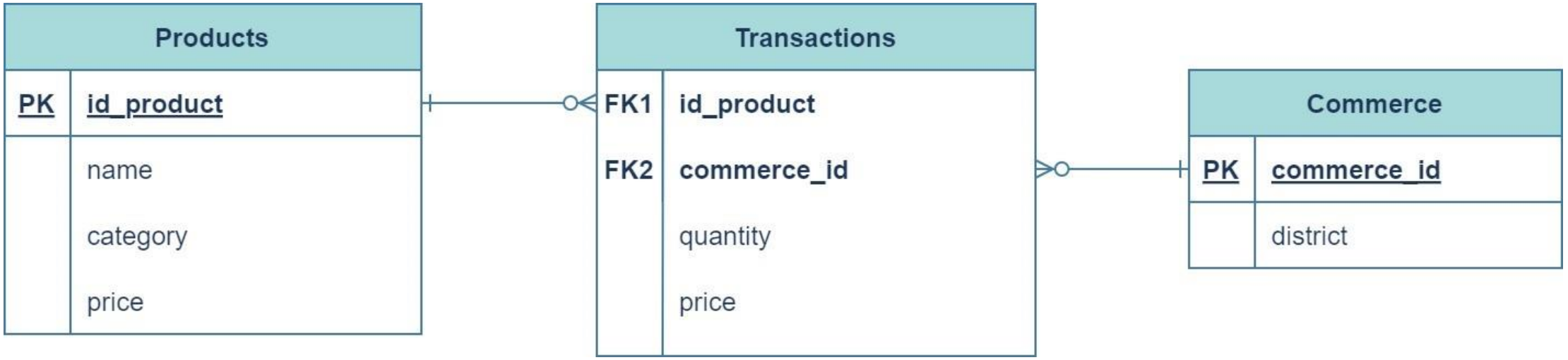
- **Productos**:

- **id_product**: ID único del producto.
- **name**: Nombre del producto.
- **category**: Categoría (Ropa, Hogar, etc).
- **price**: Precio unitario.

- **Transacciones**:

- **id_commerce**: Comercio que realizó la compra.
- **id_product**: Producto adquirido.
- **quantity**: Cantidad comprada.
- **price**: Precio total de la compra.

Modelo Relacional



Se construyó un modelo tipo estrella con la tabla **transactions** como tabla de hechos y tanto **commerce** como **products** corresponden a las tablas de dimensiones.

Hipótesis y validaciones

- Identifiqué valores duplicados en la tabla **transacciones**. Se asumió que correspondían a datos históricos, por lo que se agruparon sumando las columnas **cantidad** y **precio** para consolidar la información acerca de compras realizadas por los comercios.

- Al implementar una función inicial para realizar recomendaciones, se observó que **todos los comercios tenían al menos uno de todos los productos disponibles**, lo que invalidaba la segunda hipótes: evitar recomendar productos que ya habían sido comprados para evitar la redundancia. Esto llevó a desestimar dicha hipótesis, pues no era aplicable en este contexto.

- Se optó por **penalizar** aquellos productos. Esto significa que los productos que un comercio ya ha adquirido reciben un menor peso en su "**score**" de recomendación, sin ser eliminados por completo del sistema. Este cambio permitió ofrecer recomendaciones más relevantes y ajustadas a las preferencias reales de cada comercio, sin dejar de considerar productos que pudieran ser útiles para reposición o futuras compras.

Modelos Elegidos

Exploramos múltiples modelos para comprender cuál proporcionaba las recomendaciones más precisas y útiles:

- **Jaccard:** seleccionado por su simplicidad en términos de conjuntos, considerando productos adquiridos en común entre comercios.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

- **Coseno:** elegido para aprovechar la granularidad de los datos y medir la proximidad entre comercios basados en patrones de compra completos.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- **KNN:** se utilizó para crear recomendaciones basadas en vecindarios de comercios similares, simulando un enfoque de recomendaciones colaborativas.

Resultados por Modelos

Top 10 recomendaciones (adjusted_score)

Comuna: Macul

Comercio: 10

Jaccard

	id_product	adjusted_score	name	category	price
0	29	1.000000	Producto 29	Hogar	64
1	22	0.973587	Producto 22	Electrónica	42
2	45	0.928892	Producto 45	Hogar	33
3	17	0.897123	Producto 17	Alimentos	54
4	50	0.855023	Producto 50	Ropa	99
5	10	0.823143	Producto 10	Ropa	56
6	21	0.761777	Producto 21	Ropa	72
7	36	0.749467	Producto 36	Alimentos	55
8	30	0.731299	Producto 30	Juguetes	60
9	5	0.693181	Producto 5	Ropa	95

Coseno

	id_product	adjusted_score	category	price
28	29	1.000000	Hogar	64
44	45	0.995045	Hogar	33
21	22	0.984143	Electrónica	42
16	17	0.822597	Alimentos	54
1	2	0.801784	Electrónica	76
20	21	0.767096	Ropa	72
27	28	0.716551	Electrónica	65
9	10	0.691774	Ropa	56
10	11	0.682854	Hogar	44
29	30	0.674926	Juguetes	60

KNN

	id_product	adjusted_score	category	price
1	2	1.000000	Electrónica	76
9	10	0.880658	Ropa	56
23	24	0.841564	Juguetes	54
48	49	0.820988	Ropa	19
21	22	0.800412	Electrónica	42
28	29	0.790123	Hogar	64
33	34	0.759259	Juguetes	23
6	7	0.718107	Juguetes	80
18	19	0.681070	Alimentos	98
44	45	0.633745	Hogar	33

Evaluación del modelo

Modelo	Precisión	Recall	F1-Score
Jaccard	0.6000	1	1
Coseno	0.5000	0.8333	0.7273
KNN	0.4000	0.6667	0.6000

Jaccard: Modelo más preciso con un recall perfecto y un F1-Score alto. Lo que indica un claro sobreajuste del modelo.

Coseno: Ofrece un buen equilibrio entre precisión y recall.

KNN: Aunque es una técnica válida, parece que no es la mejor opción para este caso, dado que tiene el rendimiento más bajo en todas las métricas.

Conclusiones

- El sistema de recomendación de productos fue diseñado para comercios locales, priorizando la **personalización de productos** basados en los patrones de compra de **comercios similares** y la **popularidad** en la comuna.

- Se utilizaron tres modelos diferentes para realizar las recomendaciones: **Jaccard**, **Coseno** y **KNN**.

- El modelo basado en similitud de **coseno** fue seleccionado debido a su equilibrio entre **precisión** y **recall**, evitando el sobreajuste observado en el modelo Jaccard, que resultó en métricas excesivamente optimistas.

- La **penalización** de productos ya comprados por un comercio fue clave para evitar recomendaciones redundantes, mejorando la relevancia de las recomendaciones.

Recomendaciones

Datos temporales: Incorporar fechas de transacciones permitiría captar tendencias temporales, mejorando la calidad de las recomendaciones.

Hibridación de modelos: La combinación de diferentes modelos podría capturar aspectos más amplios de las preferencias de los comercios.

Optimización continua**: Implementar un sistema **MLOps** en línea permitiría ajustar las recomendaciones en tiempo real con datos nuevos.


yom