Generando Ofertas Focalizadas

Resumen
En este proyecto, nos enfrentamos al desafío de mejorar la experiencia de compra de los clientes de una empresa de alimentos mediante la implementación de un sistema de recomendación de productos. Utilizando datos históricos de ventas, que incluyen información sobre las compras realizadas por los clientes, los productos adquiridos y las fechas de compra, entre otros, se aplicaron algoritmos de filtrado colaborativo basado en usuarios y en items. Estos métodos comparan las preferencias de un cliente con las de otros clientes similares para recomendar productos que otros clientes con gustos similares han comprado y disfrutado, y la similitud entre los productos comprados por los clientes. El principal resultado del proyecto radica en la generación de un listado con recomendaciones puntuales para los clientes de la empresa de alimentos. La implementación de este sistema de recomendación mejora la experiencia de compra de los clientes al ofrecerles productos que realmente les interesan, aumenta las ventas al incentivar la compra de productos adicionales y permite a la empresa entender mejor las preferencias de sus clientes, lo que puede ser útil para futuras estrategias de marketing y desarrollo de productos.

Introducción

En el contexto competitivo actual de la industria de alimentos, las empresas enfrentan el desafío de maximizar la satisfacción del cliente mientras optimizan sus ventas y la gestión de inventarios. Uno de los problemas clave que surge en este entorno es cómo recomendar productos de manera eficiente a los consumidores, tomando en cuenta sus preferencias previas y patrones de compra. En este proyecto, se plantea la pregunta: ¿Cómo se puede diseñar un sistema de recomendación basado en usuarios para mejorar la experiencia de compra y aumentar las ventas en una empresa de alimentos?

El cliente potencial de este sistema de recomendación es una empresa que vende alimentos tanto a consumidores individuales como a minoristas, en un contexto organizacional donde los datos de ventas, históricos y actuales, son un recurso valioso pero subutilizado. En este contexto, la capacidad de personalizar las sugerencias de compra se convierte en una ventaja competitiva que puede incrementar la fidelización y la recurrencia de los clientes.

La literatura nacional e internacional sobre sistemas de recomendación resalta la efectividad de estos modelos en diversas áreas del marketing y ventas. Según Jannach y Adomavicius (2016), los sistemas de recomendación no solo aumentan la satisfacción del cliente, sino que también mejoran significativamente los ingresos mediante la personalización de productos. Por su parte, Zhang et al. (2019) destacan el papel de los sistemas colaborativos en la identificación de productos de alto valor para los usuarios, permitiendo a las empresas ofrecer recomendaciones personalizadas que alineen los intereses del consumidor con los objetivos comerciales. A nivel más práctico, Wang et al. (2020) exploran cómo los sistemas de recomendación son aplicados en marketing digital para optimizar la conversión de ventas, destacando su utilidad en mercados con una gran oferta de productos, como es el caso del sector alimentario.

Sin embargo, la aplicación de estos sistemas en la industria de alimentos ha sido menos explorada, particularmente en mercados emergentes. Este proyecto busca llenar ese vacío

al adaptar enfoques de recomendación para un entorno específico, donde los patrones de compra y las preferencias del cliente pueden ser fuertemente influenciados por factores estacionales y geográficos.

El sistema desarrollado permite proponer recomendaciones personalizadas para los clientes de la empresa de alimentos, ayudando a la optimización de las estrategias de mercadeo y posiblemente potenciar las ventas. No obstante, el enfoque presenta algunas limitaciones, como la necesidad de grandes volúmenes de datos históricos para obtener predicciones precisas y la complejidad de adaptar el sistema a cambios rápidos en la oferta y demanda de productos.

Finalmente, se sugiere la implementación gradual del sistema de recomendación, comenzando con un grupo selecto de productos clave y clientes, para optimizar el algoritmo y adaptarlo a las necesidades específicas del mercado, permitiendo ajustes antes de una implementación a gran escala.

Materiales y métodos

Para el proyecto, se tomaron los datos de venta de una empresa colombiana que produce alimentos. A continuación, se hace la descripción de cada variable y se comenta sobre su contenido:

1. Mes (Año/Mes):
    - Representa datos desde 202401 (enero de 2024) hasta 202407 (julio de 2024).
2. Cliente (ID de Cliente):
    - El conjunto de datos contiene 1,234,289 registros, indicando un gran número de transacciones.
3. CategoriaMarca (Código de Categoría/Marca):
    - Existen 21 categorías o marcas únicas, siendo "CM26" la más frecuente.
4. SKU:
    - Hay 224 SKUs únicos, con "M1115" como el SKU más común.
5. VtaTon (Volumen de Ventas en Toneladas):
    - El volumen medio de ventas es de 0.0307 toneladas, con un rango significativo desde -2 hasta 86.09 toneladas.
    - La presencia de valores negativos sugiere posibles devoluciones o errores de entrada de datos que necesitan más investigación.
6. VtaValor (Valor de Ventas):
    - El valor de ventas varía ampliamente de -8,957,606 a 262,380,100, con una media de 178,766.30.
    - La existencia de valores negativos indica potenciales reembolsos o ajustes.
7. Departamento (Departamento/Región):
    - Hay 27 departamentos únicos, siendo "BOGOTÁ" el más frecuente.
8. Poblacion (Población/Localidad):
    - 484 localidades únicas, con "BOGOTÁ D.C." apareciendo con más frecuencia.
9. TipologiaCliente (Tipología del Cliente):

- Existen 8 tipologías de clientes únicas, con "T16" siendo la más frecuente, representando una porción significativa de los registros.

10. DiaAtencion (Día de Atención/Servicio):
    - Hay 6 días de atención únicos, con "MA" siendo el más común.
11. CentroDespacho (Centro de Despacho) y OficinVentas (Oficina de Ventas):
    - El centro de despacho y la oficina de ventas tienen pocos códigos únicos, siendo "6038" y "6030" los más comunes, respectivamente.

Ahora, se presentan los detalles más técnicos y estadísticos del dataset:

1. Información del dataset:

```
Información del dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1234289 entries, 0 to 1234288
Data columns (total 12 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Mes               1234289 non-null  int64
 1   Cliente           1234289 non-null  int64
 2   CategoriaMarca    1234289 non-null  object
 3   SKU               1234289 non-null  object
 4   VtaTon            1234289 non-null  float64
 5   VtaValor          1234289 non-null  float64
 6   Departamento      1234289 non-null  object
 7   Poblacion         1234289 non-null  object
 8   TipologiaCliente  1234289 non-null  object
 9   DiaAtencion       1234289 non-null  object
 10  CentroDespacho    1234289 non-null  int64
 11  OficinVentas      1234289 non-null  int64
dtypes: float64(2), int64(4), object(6)
memory usage: 113.0+ MB
None
```

Se tienen 12 columnas en donde 6 son cualitativas y 6 cuantitativas, las últimas 6 se pueden separar en 4 de tipo discreto y 2 continuas.

2. Estadísticas descriptivas de las variables cuantitativas:

```
Resumen estadístico de las variables numéricas:
                Mes       Cliente         VtaTon      VtaValor  CentroDespacho  \
count  1.234289e+06  1.234289e+06  1.234289e+06  1.234289e+06    1.234289e+06
mean   2.024041e+05  3.028022e+06  3.069266e-02  1.787663e+05    6.035504e+03
std    1.990772e+00  1.538783e+04  3.111323e-01  1.302653e+06    3.005983e+00
min    2.024010e+05  3.001295e+06 -2.000000e+00 -8.957606e+06    6.001000e+03
25%    2.024020e+05  3.013169e+06  1.380000e-03  1.735713e+04    6.033000e+03
50%    2.024040e+05  3.030708e+06  4.500000e-03  4.330296e+04    6.038000e+03
75%    2.024060e+05  3.043733e+06  1.600000e-02  1.173809e+05    6.038000e+03
max    2.024070e+05  3.072785e+06  8.609400e+01  2.623801e+08    6.039000e+03

       OficinVentas
count  1.234289e+06
mean   6.031131e+03
std    1.433605e+00
min    6.030000e+03
25%    6.030000e+03
50%    6.030000e+03
75%    6.032000e+03
max    6.034000e+03
```

A pesar de que tenemos 6 variables numéricas, sólo dos de ellas tienen real sentido dentro del análisis (VtaTon y VtaValor) y este se presenta a continuación:

VtaTon (Ventas en Toneladas):

- Media: 0.0369 toneladas, pero con un mínimo de -2 toneladas, lo que sugiere posibles errores o valores negativos que podrían requerir revisión.
- El 75% de los valores son mayores a 0, lo que implica que la mayoría de las ventas son positivas.

VtaValor (Valor de Ventas):

- Media: 178,766.3, con un rango desde -8,957,660 a 262,380,000, lo cual sugiere que hay ventas con valores negativos, posiblemente devoluciones o errores.

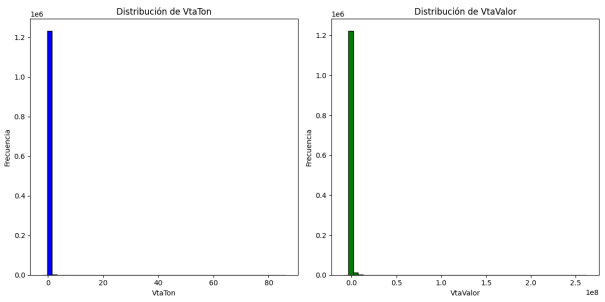
3. Valores faltantes:

Dentro de los datos no se observan datos faltantes como se muestra a continuación:

```
Valores faltantes por columna:
Mes                 0
Cliente             0
CategoriaMarca      0
SKU                 0
VtaTon              0
VtaValor            0
Departamento        0
Poblacion           0
TipologiaCliente    0
DiaAtencion         0
CentroDespacho      0
OficinVentas        0
dtype: int64
```

4. Distribución y relación de las variables numéricas:

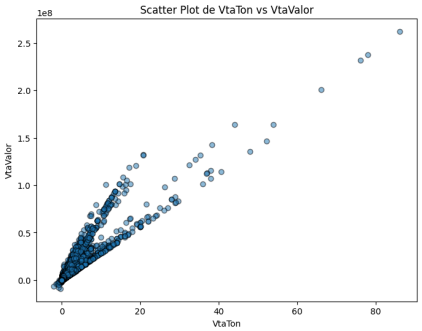
Siguiendo con el análisis de las variables, ahora se presentan la distribución de las variables VtaTon y VtaValor:



Las distribuciones de las variables VtaTon y VtaValor muestran que la mayoría de los valores están concentrados cerca de cero, con algunos valores extremos mucho más altos:

- VtaTon (Ventas en toneladas): La mayoría de las ventas en toneladas son muy pequeñas, y hay pocos casos con valores muy altos.
- VtaValor (Ventas en pesos): Similarmente, la mayoría de los valores de ventas están cerca de cero, con algunos puntos que alcanzan valores significativamente mayores.

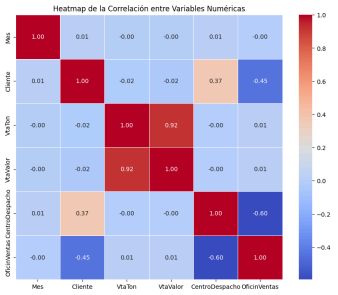
5. Gráfico de dispersión:



El gráfico de dispersión muestra la relación entre las variables VtaTon (ventas en toneladas) y VtaValor (valor de ventas). La mayoría de los puntos están agrupados cerca del origen, lo que sugiere que hay muchas observaciones con ventas bajas tanto en toneladas como en

valor. Sin embargo, también se observan algunos puntos dispersos que representan ventas mucho mayores

6. Gráfico de correlación:



- VtaTon y VtaValor tienen una alta correlación positiva (0.92), lo cual es esperado, ya que ambas métricas están relacionadas con las ventas.
- Cliente y CentroDespacho tienen una correlación moderada (0.37), sugiriendo cierta relación entre el cliente y el centro de despacho.
- CentroDespacho y OficinVentas presentan una correlación negativa significativa (-0.60), indicando que, a medida que una variable aumenta, la otra tiende a disminuir.
- No hay otras correlaciones fuertes entre las demás variables.

Finalmente, para el desarrollo del modelo, se eliminaron los registros con entradas menores a 0 en la combinación cliente y sku para luego implementar un sistema de recomendación utilizando dos algoritmos principales: SVD (Descomposición en Valores Singulares) y KNN (K-Nearest Neighbors), ambos aplicados al conjunto de datos de ventas. Se utiliza GridSearchCV para ajustar los hiperparámetros de ambos algoritmos, optimizando su rendimiento mediante la métrica RMSE. El modelo SVD descompone la matriz de interacciones usuario-producto en factores latentes para predecir la calificación de productos no comprados por los clientes, mientras que KNN encuentra los usuarios más similares para hacer recomendaciones basadas en las preferencias de otros usuarios cercanos. Ambos modelos se entrenan con los mejores parámetros encontrados, y finalmente se evalúa su precisión en un conjunto de prueba. De igual manera, pero sin usar SVD se entrenó un sistema de recomendación basado en items que proporciona recomendaciones en función de la similitud entre productos, en este caso sku. Los sistemas generan recomendaciones personalizadas para un conjunto de clientes, priorizando los productos con calificaciones estimadas más altas.

Resultados y discusión

1. Implementación de los algoritmos

Se implementaron dos enfoques: uno basado en usuarios (con SVD y KNN) y otro basado en ítems (KNN). El enfoque de SVD descompone la matriz de interacciones usuario-producto en factores latentes, mientras que KNN calcula similitudes entre usuarios o productos, dependiendo del enfoque (usuarios o ítems).

2. Resultados y comparación

- SVD fue el modelo más preciso con un RMSE de 0.6819, superando a KNN basado en usuarios (RMSE de 0.9631) y a KNN basado en ítems (RMSE de 0.8249). Esto sugiere que SVD puede captar mejor las relaciones entre usuarios y productos en este conjunto de datos.

Modelo	RMSE
SVD	0.6819
KNN (Usuarios)	0.9631
KNN (Ítems)	0.8249

3. Discusión

El algoritmo SVD fue el más efectivo, proporcionando recomendaciones más precisas. Sin embargo, KNN basado en ítems también es útil en escenarios donde las similitudes entre productos son importantes. Las limitaciones de ambos enfoques incluyen la necesidad de datos históricos y el costo computacional de SVD.

Para mejorar, podría considerarse un enfoque híbrido que combine lo mejor de ambos algoritmos.

Por último, presentamos las recomendaciones realizadas por los enfoques SVD y KNN basado en ítems para un usuario (3001295):

SVD basado en usuarios	KNN basado en ítems
Producto: H700, Calificación estimada: 3.405922570710446	Producto: Q1175, Calificación estimada: 0.07148995466283677
Producto: H950, Calificación estimada: 1.3269302267089296	Producto: H701, Calificación estimada: 0.05540062303289657
Producto: H1081, Calificación estimada: 0.8477679667307565	Producto: M1115, Calificación estimada: 0.05131149728206614
Producto: H1013, Calificación estimada: 0.6808636885724593	Producto: H913, Calificación estimada: 0.05100420283539071
Producto: M1075, Calificación estimada: 0.555646894820121	Producto: M1069, Calificación estimada: 0.048215436338084436

Conclusión

En este trabajo se implementaron y compararon dos enfoques de sistemas de recomendación para una empresa de alimentos: un sistema basado en usuarios (utilizando SVD y KNN) y uno basado en ítems (utilizando KNN). Los resultados muestran que el algoritmo SVD obtuvo el mejor rendimiento, con un RMSE de 0.6819, lo que indica una mayor precisión en las recomendaciones. En comparación, los enfoques basados en KNN presentaron un rendimiento inferior, con un RMSE de 0.9631 para KNN basado en usuarios y 0.8249 para KNN basado en ítems.

Se concluye que el enfoque basado en SVD es el más adecuado para este conjunto de datos debido a su capacidad para captar patrones latentes en las interacciones usuario-producto. No obstante, el sistema basado en ítems puede ser útil en escenarios donde las similitudes entre productos son más relevantes para las recomendaciones.

Recomendaciones: Se sugiere la implementación de SVD como método principal, complementado con un enfoque híbrido que combine la similitud entre productos y usuarios para mejorar aún más la precisión y adaptabilidad del sistema.

Bibliografía

1. Jannach, D., & Adomavicius, G. (2016). Recommendation systems: Challenges, opportunities, and research directions. *AI Magazine, 37*(3), 35-41. https://doi.org/10.1609/aimag.v37i3.2680
2. 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. https://doi.org/10.1145/3285029
3. Wang, P., Zhang, H., Wang, B., & Yuan, X. (2020). The impact of recommender systems on online marketing strategies. *Journal of Marketing Science, 18*(4), 56-72.