Primera entrega

UniChampions

En esta primera entrega se va a dará conocer las estrategias que se quieren implementar para el algoritmo de recomendación de Corona.

---

1. Introducción y Objetivos

El objetivo principal es aprovechar los datos transaccionales y de cotizaciones existentes para ofrecer a cada cliente sugerencias de productos relevantes, con el fin de mejorar su experiencia, incrementar la tasa de conversión, el valor promedio de pedido y la fidelización.

Objetivos Específicos:

- Incrementar las ventas generadas a través de recomendaciones personalizadas.
- Aumentar la satisfacción y retención del cliente mediante una experiencia de compra más relevante.

2. Alcance del Proyecto

Este proyecto se centrará exclusivamente en el segmento B2C.

- **Incluido en el Alcance:**
  - Análisis y procesamiento de las bases de datos base_1_transaccional y base_2_cotizaciones.
  - Diseño e implementación de un modelo de recomendación híbrido (combinando Filtrado Colaborativo y Basado en Contenido).
  - Generación de recomendaciones personalizadas a nivel de producto.
  - Incorporación de características de usuario (e.g., edad, municipio, zona) y de producto (e.g., categoria_macro, categoria, subcategoria, color, alineación con portafolio estratégico) en el modelo.
  - Desarrollo de estrategias para abordar el problema del "arranque en frío" (cold start) para nuevos usuarios y productos.
  - Documentación del proceso, modelos y resultados preliminares.

3. Fuentes de Datos y Relevancia

Se utilizarán las siguientes fuentes de datos proporcionadas:

1. **base_1_transaccional:**
    - **Descripción:** Historial de ventas confirmadas a clientes B2C.
    - **Relevancia:** Fuente principal de **feedback positivo explícito/implícito**. Contiene la relación cliente (id) - producto (producto) comprados, junto con metadatos valiosos de producto (categoria_macro, categoria, subcategoria, color), usuario (edad, municipio, zona), contexto (fecha, punto de venta) y negocio (alineación con portafolio estratégico). Es la base para el Filtrado Colaborativo y proporciona features para el Filtrado Basado en Contenido y modelos híbridos.
2. **base_2_cotizaciones:**
    - **Descripción:** Registro de cotizaciones realizadas a clientes B2C.
    - **Relevancia:** Captura la **intención de compra** o interés del cliente (id) en un producto (producto), incluso si no se concretó la venta. El campo estado_cotizacion es crucial para ponderar este interés ('Pendiente' como señal fuerte de interés actual, 'Ganada' como confirmación). Permite enriquecer el perfil del usuario y abordar el cold start si un usuario ha cotizado pero no comprado. Comparte la misma estructura de metadatos de producto que base_1. Se asume que el id es consistente entre base_1 y base_2 para un mismo cliente.
3. **base_3_transaccional_b2b:**
    - **Descripción:** Historial de ventas a clientes B2B.
    - **Relevancia: Nula para el objetivo B2C.** Contiene identificadores (id_b2b) y categorías diferentes. **Esta base de datos será explícitamente excluida** del análisis y modelado para evitar introducir ruido y sesgos irrelevantes en las recomendaciones B2C.

4. Metodología Propuesta (Paso a Paso)

Se seguirá un enfoque iterativo y basado en datos, dividido en las siguientes fases principales:

4.1. Fase 1: Comprensión y Exploración de Datos (EDA - Exploratory Data Analysis)

- **Objetivo:** Entender a fondo los datos, identificar patrones, inconsistencias y características relevantes.
- **Tareas:**
    - Cargar los datasets (base_1, base_2) en un entorno de análisis (e.g., Python con Pandas).

- Realizar análisis estadísticos descriptivos (distribuciones de edad, cantidad, valor, frecuencia de categorías, etc.).
- Verificar la calidad de los datos: identificar valores faltantes, duplicados, formatos inconsistentes, rangos anómalos (edad, precio).
- Analizar la distribución de interacciones: ítems más populares, usuarios más activos, dispersión de la matriz usuario-ítem.
- Visualizar relaciones: correlaciones, tendencias temporales (fecha), patrones geográficos (municipio, zona).
- Confirmar la consistencia del id entre base_1 y base_2.
- Documentar hallazgos clave que informarán las siguientes fases.

4.2. Fase 2: Preprocesamiento y Consolidación de Datos

- **Objetivo:** Preparar los datos crudos en un formato limpio y estructurado, adecuado para el modelado.
- **Tareas:**
  - **Limpieza:** Imputar o eliminar valores faltantes (según estrategia definida), corregir inconsistencias, estandarizar nombres/categorías.
  - **Consolidación:** Crear una tabla/dataframe unificado de interacciones B2C combinando base_1 (compras) y base_2 (cotizaciones). Columnas clave: user_id, item_id, interaction_type (compra, cotizacion_pendiente, etc.), timestamp, weight (asignar pesos diferentes a compras vs. cotizaciones, e.g., compra=1.0, cotización_pendiente=0.5), cantidad, valor.
  - **Mapeo de IDs:** Crear identificadores numéricos únicos para cada user_id y item_id si los modelos lo requieren, guardando los mapeos.
  - **Codificación:** Convertir variables categóricas (categoria_macro, categoria, subcategoria, color, municipio, zona) a representación numérica (e.g., One-Hot Encoding, Label Encoding, o preparación para Embeddings).
  - **Tratamiento de Fechas:** Convertir fecha, fecha_creacion a formato datetime. Extraer componentes útiles (mes, día de la semana) si se considera relevante.

4.3. Fase 3: Ingeniería de Características (Feature Engineering)

- **Objetivo:** Crear nuevas características informativas a partir de los datos existentes para mejorar el rendimiento del modelo.
- **Tareas:**

- **Features de Interacción Usuario-Ítem:** (e.g., frecuencia de compra/cotización, tiempo desde la última interacción).
- **Features de Usuario:** Agregar características del usuario (e.g., edad normalizada, zona/municipio codificados). Potencialmente, features agregadas (gasto total/promedio, categorías preferidas).
- **Features de Ítem:** Agregar características del producto (categorías codificadas, color, alineación con portafolio estratégico como feature numérica/categórica).
- **Features Contextuales:** Considerar hora del día (si disponible y relevante), día de la semana extraído de fecha.

4.4. Fase 4: Selección y Diseño del Modelo de Recomendación

- **Objetivo:** Elegir y diseñar la arquitectura del modelo más adecuada para los datos y objetivos.
- **Justificación del Enfoque Híbrido:** Dados los datos ricos (interacciones + metadata de usuario/ítem), un enfoque híbrido es el más prometedor, combinando las fortalezas del Filtrado Colaborativo (patrones de comportamiento) y el Filtrado Basado en Contenido (similitud de ítems/usuarios, manejo de cold start).
- **Modelos a Considerar/Evaluar:**
  - **Filtrado Colaborativo (Item-Based):** Calcular similitud entre productos basada en qué usuarios los compraron/cotizaron juntos. Generar recomendaciones basadas en ítems similares a los que el usuario interactuó.
  - **Factorización de Matrices (e.g., SVD, ALS):** Descomponer la matriz de interacción usuario-ítem en factores latentes para predecir preferencias. Adecuado para feedback implícito (usando pesos definidos en Fase 2).
  - **(Opcional/Futuro): Modelos de Deep Learning.** Como Wide & Deep o Neural Collaborative Filtering extendidos con features. Requieren más datos y esfuerzo, pero pueden ofrecer mayor rendimiento. Se considerarán en iteraciones posteriores si es necesario.
  - **Idea:** También se busca evaluar la idea de tener diferentes modelos, uno puede ser para los nuevos clientes y otro para los antiguos.

4.5. Fase 5: Entrenamiento y Evaluación Offline

- **Objetivo:** Entrenar los modelos seleccionados y evaluar su rendimiento utilizando datos históricos antes de considerar un despliegue.
- **Tareas:**

- **División de Datos:** Separar el conjunto de datos consolidado en conjuntos de entrenamiento, validación y prueba. **Es crucial usar una división temporal:** entrenar con datos hasta una fecha T1, validar (para ajuste de hiperparámetros) en el período T1-T2, y testear el rendimiento final en datos posteriores a T2. Esto simula un escenario real.
- **Entrenamiento:** Entrenar los modelos seleccionados (Baseline, CF, LightFM) sobre el conjunto de entrenamiento.
- **Ajuste de Hiperparámetros:** Utilizar el conjunto de validación para encontrar la mejor configuración para cada modelo (e.g., número de factores, tasa de aprendizaje, regularización).
- **Evaluación:** Evaluar los modelos finales sobre el conjunto de prueba utilizando métricas de ranking apropiadas:
  - **NDCG@k (Normalized Discounted Cumulative Gain):** Mide la calidad del ranking (relevancia y posición). Clave.
  - **Precision@k:** Proporción de recomendaciones relevantes entre las k primeras.
  - **Recall@k:** Proporción de ítems relevantes consumidos por el usuario que fueron recomendados entre los k primeros.
  - **MAP@k (Mean Average Precision):** Promedio de Precision@k considerando el orden.
- **Comparación:** Comparar el rendimiento de los diferentes modelos contra el baseline y entre sí.

4.6. Fase 6: Estrategias Específicas

- **Objetivo:** Abordar desafíos conocidos de los sistemas de recomendación.
- **Tareas:**
  - **Manejo del Cold Start (Nuevos Usuarios):** Definir estrategia: recomendar ítems populares (global, por zona, por edad), usar cotizaciones (base_2) si existen, o basarse inicialmente en features demográficas/geográficas si el modelo lo permite (LightFM).
  - **Manejo del Cold Start (Nuevos Productos):** Si el producto tiene metadatos (categoria, color), usar un enfoque basado en contenido para encontrar usuarios interesados en productos similares. Si no, no se podrá recomendar hasta obtener interacciones.

- **Promoción de Productos Estratégicos:** Investigar cómo usar el campo alineación con portafolio estratégico para "impulsar" o dar mayor visibilidad a ciertos productos en la lista final de recomendaciones (e.g., mediante re-ranking o incorporándolo como feature con peso en el modelo).

4.7. Fase 7: Consideraciones para Despliegue Futuro (Planificación)

- **Objetivo:** Pensar en cómo el modelo se integrará y operará en un entorno real.
- **Tareas (Conceptuales en esta fase):**
  - **Formato de Entrega:** ¿Cómo se entregarán las recomendaciones? (e.g., API REST que recibe user_id y devuelve lista de item_id recomendados).
  - **Frecuencia de Actualización:** ¿Con qué frecuencia se reentrenará el modelo? (Diario, semanal). ¿Se pueden actualizar solo las recomendaciones (inferencia) más frecuentemente?
  - **Infraestructura:** ¿Qué recursos computacionales se necesitarán para entrenamiento e inferencia?
  - **Monitorización:** ¿Qué métricas se seguirán en producción? (Además de métricas de negocio, latencia de API, tasa de error).
  - **A/B Testing:** Planificar cómo se implementaría un sistema de A/B testing para comparar el modelo con un grupo de control o versiones anteriores.

5. Estrategia de Evaluación del Éxito

El éxito del proyecto se medirá en dos niveles:

- **Evaluación Offline (Durante el Desarrollo):**
  - Utilizando las métricas de ranking definidas (NDCG@k, Precision@k, Recall@k, MAP@k) sobre el conjunto de prueba temporal.
  - El objetivo es superar significativamente al modelo baseline y demostrar la efectividad del enfoque híbrido propuesto.

6. Información Adicional Requerida (Pendiente)

- Confirmación final sobre la consistencia y método de enlace del campo id entre base_1_transaccional y base_2_cotizaciones. En pocas palabras, ¿Cómo puedo identificar al cliente que hizo la cotización en la tabla base_2_cotizaciones para poder unirlo con los id's de los clientes de la tabla base_1_transaccional?
- ¿En los datos B2B no me queda claro si son las ventas realizadas a otras empresas o si son las ventas de las empresas que compran nuestro producto y después lo

venden a los clientes ellos mismo? ¿En dado caso que fuera la segunda entonces los identificadores de id_b2b de esta tabla B2B serian los mismo que los de id en B2C?

- En el documento se dice que el recomendador será una herramienta clave para los asesores de venta en los puntos físicos, lo que no nos queda claro es si el recomendador solo se usara para los asesores de venta física o también en la plataforma online como recomendaciones de compra. Por otro lado, en cualquiera de los dos casos, también nos preguntamos ¿si toca hacer una interfaz de usuario para los asesores de venta donde pongan el id del comprador y este le de las recomendaciones? Y en caso tal de que también el recomendador sea para plataformas online, ¿toca hacer interfaz de usuario para esta también?
- ¿Por último, se sabe como se van a evaluar estos modelos, por ejemplo, van a probarlo con datos históricos o tienen algunas métricas cuantitativas que nos puedan dar como para dar una idea de cómo se van a evaluar los modelos?