**Caso de Negocio:** Cervecería y Maltería Quilmes cuenta con una aplicación B2B llamada BEES, donde comercializa sus productos a distintas empresas. Se realizó un estudio de satisfacción del cliente mediante encuestas con la aplicación y determinaron que uno de los puntos clave de mejora era desarrollar un nuevo módulo dentro de la aplicación que le permita a los clientes encontrar los productos más propensos a adquirir en su próximo pedido mejorando su experiencia y reduciendo el tiempo destinado a efectuar cada compra.

**Solución:** El equipo de ciencia de datos de la compañía le encomiendo a usted desarrollar un modelo de Machine Learning que permita discernir que productos comprará cada uno de sus clientes en su próximo pedido para ser incluidos en el nuevo módulo de la aplicación.

**A** **considerar:** Existe en la compañía un equipo de “Producto” que va a tomar como input el output de su modelo. El output de su modelo debe ser especificado para cada cliente y va a tener una frecuencia diaria. Este output se va a integrar en la arquitectura de datos de la app. El equipo de ingeniería se encarga de esto mediante pipelines. De esta manera el equipo de producto, va a tomar este input, le va a asignar un diseño y lo va a visualizar dentro de la aplicación como un nuevo módulo.

**Datasets:**

Para el caso de estudio, se proveen los siguientes conjuntos de datos:

* trasacciones.csv
* atributos.csv

abreviaciones: CMQ à Cervecería y Maltería Quilmes.

trasacciones.csv

Este dataset contiene información transaccional de las compras realizadas por los clientes en la aplicación BEES entre finales de mayo y agosto de 2022. A continuación, se detalla el diccionario de variables de esta tabla:

* ACCOUNT\_ID: código de identificación del cliente.
* SKU\_ID: código de identificación de un producto.
* INVOICE\_DATE: fecha en que se realizó la transacción en formato ‘yyyyMMdd’
* ORDER\_ID: código identificador de un pedido.
* ITEMS\_PHYS\_CASES: cantidad de bultos que compra el cliente de un determinado SKU\_ID

Así se ven algunos registros de esta tabla:



atributos.csv

contiene atributos asociados a clientes. Las variables incluidas son:

* POC: código de identificación del cliente, semánticamente, significa lo mismo que el campo “ACCOUNT\_ID” en el dataset expuesto anteriormente.
* BussinessSegment: perfil digital del cliente, contiene los siguientes atributos

|  |  |
| --- | --- |
| Variable | referencia |
| PowerUsage | Indica que el cliente tiene un nivel digitalización alto, usa todo el ecosistema de la aplicación BEES |
| HighUsage | Indica que el cliente tiene un nivel digitalización alto, pero no usa todo el ecosistema de la aplicación BEES. |
| MediumUsage | Indica que tiene un nivel de uso intermedio de la aplicación en comparación con otros clientes. |
| MinimalUsage | Nivel de uso de la aplicación más bajo o nulo. |

* totalVolumen: volumen en hectolitros adquirido por el cliente en los últimos tres meses cerrados del periodo expuesto, es decir, de junio a agosto.
* SkuDistintosPromediosXOrden: cantidad de productos distintos promedio adquiridos en ordenes en el periodo expuesto, es decir, de junio a agosto.
* SkuDistintosToTales: cantidad de productos distintos adquiridos en el periodo expuesto, es decir, de junio a agosto.
* Concentración: estimado de concentración de otros negocios de diferente índole a 150 metros cuadrados del punto de venta en cuestión, por ejemplo, un kiosco que venden Quilmes, si tiene carnicerías, tienda de ropa, panaderías ect, tiene una concentración “Alta” a su alrededor. Pueden ser Alto, Medio, Bajo o S/D que significa sin datos.
* nse: nivel socioeconómico de habitantes alrededor del punto de venta. Puede ser Alto, Medio, Bajo o S/D que significa sin datos.
* segmentoUnico: hace referencia a la capacidad de compra de cliente en cuestión.

|  |  |
| --- | --- |
| Variable | referencia |
| 1.Inactivos | Cliente con baja capacidad de compra |
| 2.Masivos | Cliente con media capacidad de compra |
| 3.Potenciales | Cliente con alta capacidad de compra |
| 4.Activos | Clientes con capacidad de compra intermedia y que, además, adquiere productos comercializados por CMQ que pertenecen a terceros. |

* Canal: canal de marketing al que pertenece el cliente, los valores que puede tener son: COMIDA, Tradicional, BEBIDA, Mayorista, Kioscos/Maxikioscos, OTROS REF; Instituciones, ENTRETENIMIENTO, Autoservicio, KA Minoristas,

Bar/Restaurant.

Así se ven algunos registros de esta tabla:



**Evaluación**

1. “EDA”. La idea de este punto es que integre los dataset compartidos, se plantee una hipótesis como por ejemplo “que variables me ayudarían a realizar predicciones” y como el Feature engineering podría mejorar la performance del modelo. Desarrolle algunas visualizaciones que fundamenten dichas hipótesis.
2. “Data wrangling y modelado” desarrolle el preprocesing y modelado, justifique asunciones que tome durante el proceso, y el o los modelos elegidos.
3. Evalué el o los modelos desarrollados en distintos periodos, además, justifique las métricas que selecciono y el periodo de tiempo definido para la misma.
4. “Output”: Defina cual es el output de su modelo. Que columnas va a tener. Para esta parte, va a tener que tomar 2 definiciones: a) ¿Cuántas recomendaciones de productos va a hacer para cada cliente? ¿Cuál va a ser el criterio para definir esta cantidad?

**Entregables**

Como requisito se debe entregar:

1. las notebooks con los desarrollos que realizo para responder los puntos de la evaluación.