**Archivo de Documentación**

**Tabla de Contenido**

1. Decisiones tomadas en el preprocesamiento y el modelado
2. Explicaciones sobre las visualizaciones creadas y los insights obtenidos
3. Discusión de posibles mejoras o limitaciones del modelo

**Descripción**

**Decisiones tomadas en el preprocesamiento y el modelado**

En la etapa de preprocesamiento, se llevaron a cabo diversas acciones para limpiar y preparar los datos. Se eliminaron los valores nulos utilizando una estrategia de imputación con la media o mediana, dependiendo de la distribución de los datos. También se eliminaron algunas filas (drop) que contenían valores irrelevantes o erróneos.

Se trataron los outliers utilizando la variable 'cantidad' para evitar el impacto negativo en el rendimiento del modelo.

Se cambiaron los tipos de variables según la relevancia y el análisis de los datos.

El dataset se filtró para separar la atención directa e indirecta, y se agrupó de la siguiente manera para calcular la probabilidad de compra semanal por cliente y producto:

Esta función basándose en el historial de compras. Primero, agrupa los datos por cliente, producto, y otras variables como la semana, el canal de distribución y el sector, sumando las cantidades compradas. Luego, calcula el total de compras por cliente y semana, considerando las mismas variables agrupadas. Este total se utiliza para calcular la probabilidad de que un cliente compre un producto específico, dividiendo la cantidad comprada de ese producto entre el total de compras del cliente en esa semana. Finalmente, la función devuelve un DataFrame ordenado que contiene esta probabilidad para cada cliente y producto. Además, se utilizó esta probabilidad para definir una variable target que indica si la probabilidad de compra es mayor a 0.5 (valor 1) o no (valor 0), facilitando la clasificación en el modelo.

Se probaron diferentes modelos, incluidos Árboles de Decisión y Random Forest. El modelo que mejor rendimiento mostró fue Random Forest, con una precisión del 70% de f1 score al ser esta una problemática de clasificacion. La métrica de f1 score entre precisión y recall, proporcionando un equilibrio entre ambas métricas y recall Mide la capacidad del modelo para identificar todos los casos positivos, (es decir, los clientes que efectivamente compraron el producto). Esto es crucial cuando queremos evitar perder oportunidades de venta, ya que preferimos un modelo que capture la mayor cantidad posible de clientes que tienen alta probabilidad de compra.

Además, el modelo fue probado con datos filtrados de los clientes 330740, 331160 y 336750.

**Explicaciones sobre las visualizaciones creadas y los insights obtenidos**

* En la etapa de preprocesamiento, se llevaron a cabo diversas acciones para limpiar y preparar los datos. Se eliminaron los valores nulos utilizando una estrategia de imputación con la media o mediana, dependiendo de la distribución de los datos. Además, se normalizaron algunas variables continuas para mejorar el rendimiento del modelo. También se realizaron codificaciones de variables categóricas utilizando enfoques como One-Hot Encoding según la relevancia de cada variable para el modelo.
* **DIST\_CHANNEL\_NAME**: La mayoría de las transacciones se concentran en los canales de Autoservicios y Tradicional, con una mayor frecuencia en el canal Tradicional.
* **SUB\_DIST\_CHANNEL\_NAME**: Las subcategorías más comunes incluyen Tiendas y Distribuidores, con una distribución notable en el segmento de Tiendas.
* **TIPOL\_TRANS\_NAME**: Se observó una alta frecuencia en Tipos de Transacción como Tienda Mixta e Hipermercado, mientras que otras categorías tienen una participación considerablemente menor.
* **SECTOR\_NAME**: El sector con mayor volumen de transacciones fue Chocolates, seguido de Galletas, lo cual destaca una tendencia importante hacia estos productos.
* **BRAND\_NAME**: Las marcas con mayor frecuencia fueron Noel y Ducales, indicando su predominancia en las transacciones analizadas.
* **CATEGORY\_NAME**: Las categorías más representadas fueron Galletas y Cereales, lo cual puede sugerir una mayor preferencia por estos tipos de productos.

**Conclusión**: Las visualizaciones creadas permitieron identificar patrones clave en los datos, como la predominancia de ciertos canales de distribución y categorías de productos. Estos insights ayudan a enfocar las estrategias comerciales en las áreas de mayor impacto y a tomar decisiones informadas sobre la preparación y transformación de los datos.

* El modelo podría beneficiarse de un mayor volumen de datos para reducir la varianza y mejorar la capacidad de generalización. Además, algunas de las variables categóricas tienen muchas categorías únicas, lo cual podría causar sobreajuste. Se podría considerar una técnica de reducción de categorías o la aplicación de un enfoque más robusto para el tratamiento de variables categóricas. Además, se podrían explorar métodos adicionales de regularización para mejorar la robustez del modelo.

**Discusión de posibles mejoras o limitaciones del modelo**

**Categorías Únicas en Variables Categóricas**

Algunas variables categóricas tienen muchas categorías únicas, lo cual podría causar sobreajuste. Se podría considerar una técnica de reducción de categorías o la aplicación de un enfoque más robusto para el tratamiento de variables categóricas, como agrupar categorías menos frecuentes.

**Evaluación de Modelos Alternativos:**

valuar otros modelos de clasificación podría ser útil para encontrar una mejor arquitectura que se adapte a las características específicas del conjunto de datos.

**Feature Engineering Adicional:**

Se podrían crear nuevas características a partir de las existentes para proporcionar más información al modelo y mejorar su capacidad predictiva. Esto podría incluir interacciones entre variables o la transformación de características numéricas en categóricas relevantes.

**Conclusión**: Para mejorar el rendimiento y la robustez del modelo, es fundamental considerar un enfoque, reducción de la complejidad de las variables categóricas. Evaluar diferentes arquitecturas de modelos, también feature engineering permitirá mejorar la capacidad predictiva.