Tarea 1 - Problem Set

Christofer Palominos Navarrete[1] and José Luis Gajardo Angel[1]

[1]Universidad Diego Portales , Facultad de Administración y Economía / Facultad de Ingeniería y Ciencias

22 de enero de 2026

Análisis de la Data

Limpieza de Caracteres No Numéricos

Al cargar la información y generar aritmética con ella se producía un `TypeError`. Este error ocurría porque columnas como `precio_unitario_venta` o `costo_unitario` contenían símbolos de moneda ($), comas (,) o espacios.

Solución: Se implementó la función `clean_numeric` con expresiones regulares para forzar la conversión a formato numérico (*floats*).

Normalización de Categorías y IDs

Surgían inconsistencias por errores de escritura (ej. "Sauld" por "Salud"). Se aplicó `.str.strip().str.title()` para unificar criterios.

Integración con el Maestro de Productos

Se utilizó el archivo `maestro_productos.csv` como fuente de verdad mediante un *merge* por `product_id`, asegurando que los costos faltantes en las transacciones se completen correctamente.

---

1. Cálculo de KPI's

1.1. Cero Censurado (Out-Of-Stock) y Sustituciones

El Cero Censurado ocurre cuando la demanda real no se registra por falta de stock. Esto genera una subestimación de la demanda futura. Las sustituciones inflan artificialmente las ventas de productos secundarios.

Medidas de Ingeniería:

- **Imputación de Demanda:** Estimar ventas perdidas basadas en el histórico.
- **Flags de Ruptura:** Variable binaria para que los modelos ignoren días de stock cero.

1.2. GMROI (Gross Margin Return on Investment)

Indica cuánta utilidad bruta genera cada peso invertido en inventario.

$$\text{GMROI} = \frac{\text{Margen Bruto}}{\text{Costo Promedio del Inventario}}$$

Un margen bajo (ej. Panadería) puede tener mayor GMROI que uno de lujo (ej. Joyería) si su **rotación** es significativamente más alta.

1.3. Porcentaje de Markdown

Mide la pérdida de ingreso potencial por rebajas. Un **Markdown >20 %** con GMROI alto suele indicar sobre-stock o una estrategia de *pricing* inicial errónea.

	categoria	GMROI	Markdown_Pct
137	Hogar	94.049948	6.983152
229	Relojería	86.077190	4.746089
58	Carnes	84.140954	5.391397
151	Informática	81.553854	6.693066
196	Muebles	81.403035	3.758732
240	Ropa Hombre	80.933400	5.923863
86	Deportes	80.854814	9.166626
20	Automotriz	79.969620	5.207047
216	Pequeños Electrodomésticos	79.892662	4.808126
180	Lácteos	79.524157	6.643881

Figura 1: Visualización de KPIs de inventario.

2. Web Scraping

2.1. Cuadro de la Data Obtenida

Métrica	Casas	Deptos
Propiedades filtradas (#)	670	489
Mediana precio (UF)	8570	5750
Promedio precio (UF)	9776	6075
Precio por m2 (UF/m2)	62.43	66.06

Figura 2: Visualización de cuadro de resultados web scraping.

2.2. Desafíos Técnicos, Éticos y Medidas de Producción

Durante el desarrollo, se enfrentaron desafíos que ilustran la complejidad de la inteligencia de datos:

1. Desafíos Técnicos (Evasión y Bloqueos):

- **WAF:** Los errores de *Timeout* se debieron a firewalls que detectan el *handshake* de Playwright analizando la propiedad `navigator.webdriver`.
- **Intercepción de Clics:** Banners de cookies y tutoriales (*coach marks*) bloquearon físicamente el botón "Siguiente".
- **Carga Perezosa (Lazy Loading):** Datos como los m^2 no cargan hasta hacer *scroll*, obligando a simular comportamiento humano.
- **Testing en maquina para verificar compatibilidad:** El proceso se genero en una Macbook Pro con Sequoia 15.7.3, cargado a Git se descarga para testearlo en otra maquina con Windows 11 arrojando errores, por lo que se generan los siguientes cambios para estandarizar el proceso:
  - Migración a sync_api: Se reemplazaron los await inernos por llamadas directas.
  - page_wait_for_timeout:Se reemplazaron las llamadas asyncio.sleep dentro de la logica del navegador.
  - ThreadPoolExecutor: Lo utilizamos para encapsular el proceso para que VS Code no detecte que estamos bloquenado el hilo principal, puesto que Jupyter utiliza internamente un bucle para gestionar la comunicacion con el navegador, con esto movemos toda la logica de PlayGround a un hilo separado donde no hay bucles d eventos preexistentes, permitiendo que la version sincronica funcione sin inerferencias.

2. Desafíos Éticos (Respeto y Privacidad):

- **Carga del Servidor:** Se implementaron pausas (`asyncio.sleep`) para no saturar el sitio.

- **Veracidad y Ruido:** Se filtraron *outliers* (precios en Pesos vs UF) para evitar métricas engañosas.

3. Medidas de Producción:

- **Proxies Residenciales:** Rotación de IPs para evitar bloqueos geográficos.
- **Resolución de CAPTCHAs:** Integración de servicios de IA para retos visuales.
- **Arquitectura Limpia:** Proceso de deduplicación por ID antes del análisis.

3. Estrategia de retención y desafiliación / churn

Que estrategía se utilizara

Utilizaré un Random Forest Classifier (Bosque Aleatorio), ya que es un modelo robusto que generalmente ofrece buenos resultados en problemas de clasificación mixtos sin necesidad de un preprocesamiento excesivo.

3.1. Entrena un modelo de clasificación para predecir churn.

Primero, cargamos los datos, separamos las variables predictoras (X) de la variable objetivo (y) y entrenamos el modelo.

3.1.1. Resultados del Entrenamiento (Interpretación):

Variable Crítica: La satisfaccion es, por mucho, la variable más importante (72.3 %). Esto indica que la percepción subjetiva del cliente es el principal motor de fuga.

Soporte: Con un 17 %, la cantidad de veces que un cliente contacta a soporte es el segundo factor relevante.

Desempeño: El modelo tiene una precisión del 83 % para la clase de churn. Aunque el recall parece bajo (31 %), el análisis de deciles demuestra que el modelo es excelente ordenando el riesgo.

3.2. Realiza un análisis de lift y comenta los resultados.

Poder del Decil 1: El Lift del primer decil es de 4.71. Esto significa que en el 10 % de clientes con mayor probabilidad de fuga según el modelo, hay 4.7 veces más churners que si eligieras clientes al azar.

Eficiencia de la Estrategia: Al observar la Ganancia Acumulada, vemos que contactando solo al 20 % de la base (los dos primeros deciles), capturamos el 76.8 % de todas las fugas reales.

3.2.1. Conclusión de Negocio

El modelo permite optimizar los recursos de retención de manera masiva. En lugar de contactar a todos los clientes, la empresa puede enfocarse solo en el 20 % con mayor riesgo y lograría evitar casi el 77 % de las pérdidas de clientes.

4. CLTV

4.1. Comparación de los modelos

Regresión Lineal: Verás que el R^2 es muy alto (cercano a 0.89). Esto es porque la regresión lineal es un modelo "espejo": simplemente replica la multiplicación de (Frecuencia * Gasto) que ya existe en los datos. No "prende.[el] comportamiento futuro, solo describe el pasado.

Gamma-Gamma: Este modelo no intenta ajustar los datos pasados a la perfección. Su objetivo es estimar el valor intrínseco del cliente. En el gráfico de distribución, verás que Gamma-Gamma suele ser más "estrecho" porque elimina el ruido de los gastos extremos.

4.2. Tratamiento de Outliers y Robustez

El problema del Lineal: Si un cliente gastó $1,500 una sola vez, la Regresión Lineal asume que ese es su nivel real y le asigna un CLTV altísimo. Si ese gasto fue una anomalía (outlier), la regresión está cometiendo un error costoso para la empresa.

La solución Gamma-Gamma (Shrinkage): Este modelo aplica una técnica llamada "contracción bayesiana".

Si un cliente tiene poca frecuencia (ej. 1 o 2 compras) pero un gasto muy alto, el modelo "desconfía" de ese dato y ajusta la predicción hacia la media de la población.

¿Por qué es más robusto? Porque evita que casos aislados (ruido) distorsionen las predicciones. Protege al negocio de considerar "VIP" a alguien que solo tuvo un gasto grande por casualidad, enfocando los esfuerzos de retención en clientes cuyo valor es consistente en el tiempo.

5. Inferencia Causal

Pasos a realizar

Para este análisis de Inferencia Causal, utilizaremos el dataset data_inferencia_causal.csv. El objetivo es medir el CATE (Conditional Average Treatment Effect), que nos dice cómo varía el impacto de la publicidad según las características del cliente (como la edad o el ingreso).

5.1. Estimación del CATE: S-Learner vs. T-Learner

Implementaremos dos "Meta-Learners" utilizando modelos de Regresión Lineal (o Random Forest) para estimar el efecto individual.

5.1.1. Interpretación de los modelos

S-Learner: Es más eficiente con los datos porque usa toda la muestra, pero puede "ignorar" el tratamiento si las otras variables son muy fuertes.

T-Learner: Es más flexible para detectar efectos complejos porque permite que el impacto de la edad o el ingreso sea totalmente diferente para los que vieron el anuncio vs. los que no. Es el estándar para detectar Uplift.

5.2. Asignación Presupuestaria y Estrategia

Si el modelo detecta que el efecto de la publicidad tiene una correlación negativa con el ingreso, esto significa que:

A medida que el ingreso del cliente aumenta, la efectividad de la publicidad disminuye.

Decisión Estratégica para el Gerente de Marketing

Reasignación de Targeting (Eficiencia): El Gerente debería mover el presupuesto de los segmentos de altos ingresos hacia los de bajos ingresos.

Los clientes de ingresos altos probablemente comprarán de todos modos (efecto orgánico) o son indiferentes a la publicidad. Invertir en ellos es un "desperdicio"de presupuesto.

Los clientes de ingresos bajos muestran un mayor Uplift (reaccionan más al anuncio), por lo que cada dólar invertido allí genera más ventas incrementales.

Optimización del ROAS (Return on Ad Spend): Al enfocarse en quienes tienen un CATE más alto, el retorno de inversión publicitaria mejora significativamente. La estrategia no es buscar a quien "gasta más", sino a quien "gasta más gracias al anuncio".

Personalización del Mensaje: Si se decide mantener presencia en el segmento de altos ingresos, se debe cambiar radicalmente la creatividad. Una correlación negativa sugiere que el mensaje actual no resuena con ese grupo.

Conclusión

La inferencia causal permite pasar de un marketing de "alcance" (llegar a todos) a un marketing de "persuasión"(llegar solo a quienes el anuncio realmente logra cambiar su comportamiento).