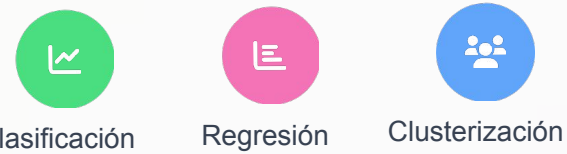
Proyecto de Aprendizaje Automático para Supermercados

Optimización de Estrategias de Fidelización y Marketing

Desarrollado por:

Irene Valle, Nicolás Febrero, Sergio Arca y Alex Caride



Fecha de informe: 09/12/2025

Objetivos del Proyecto

Desarrollo de soluciones de machine learning para una cadena de supermercados

Clasificación

Predecir la propensión de los clientes a aceptar campañas de marketing.

- Identificar clientes potencialmente interesados
- Optimizar estrategias de marketing
- Reducir costes de adquisición

Regresión

Estimar el gasto anual total de cada cliente.

- Previsión de gastos económicos
- Planificación estratégica
- Análisis de comportamiento de compra

Clusterización

Agrupar clientes para definir perfiles y segmentar el mercado.

- Identificación de segmentos de clientes
- Perfiles de consumo detallados
- Estrategias de fidelización personalizadas

Código Clasificación

Desarrollo de soluciones de machine learning para una cadena de supermercados

Clasificación

```python
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC


modelos_clasificacion = [
    ("Regresión Logística", LogisticRegression(class_weight='balanced',max_iter=1000, random_state=42)),
    ("Árbol de Decisión", DecisionTreeClassifier(class_weight='balanced', random_state=42)),
    ("SVC", SVC(probability=True, random_state=42))
]
```

- Se usa para la clasificación:
  - Regresión Logística
  - Árbol de Decisión
  - SVC

Código Regresión

Desarrollo de soluciones de machine learning para una cadena de supermercados

Regresión

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor

modelos = [
    ("Regresión Lineal", LinearRegression()),
    ("Árbol de Decisión", DecisionTreeRegressor()),
    ("SVR", SVR()),
]
```

- Se usa para la regresión:
  - Regresión Lineal
  - Árbol de Decisión
  - SVR

Código Clusterización

Desarrollo de soluciones de machine learning para una cadena de supermercados

Clusterización

```
kmeans_sklearn = KMeans(n_clusters=4, random_state=RANDOM_STATE, init="k-means++", n_init=50, max_iter=500)
kmeans_sklearn.fit(df.values)
labels_sklearn = kmeans_sklearn.predict(df.values)
```

- Para la clusterización:
  - K = 4
  - 500 iteraciones

Análisis Exploratorio de Datos

Tamaño del Dataset

1989 filas y 38 columnas

Datos Sensibles

Se identificó la presencia de:

- Nombres y apellidos
- DNI
- Números de teléfono

Valores Ausentes

Aproximadamente el 4-5% de las entradas presentaban información ausente:

- Ingresos
- Año de nacimiento

Vista Previa

nombre	estado_civil	educacion	ingresos
Juan Pérez	Casado	Universitario	-
María García	-	Secundario	35000

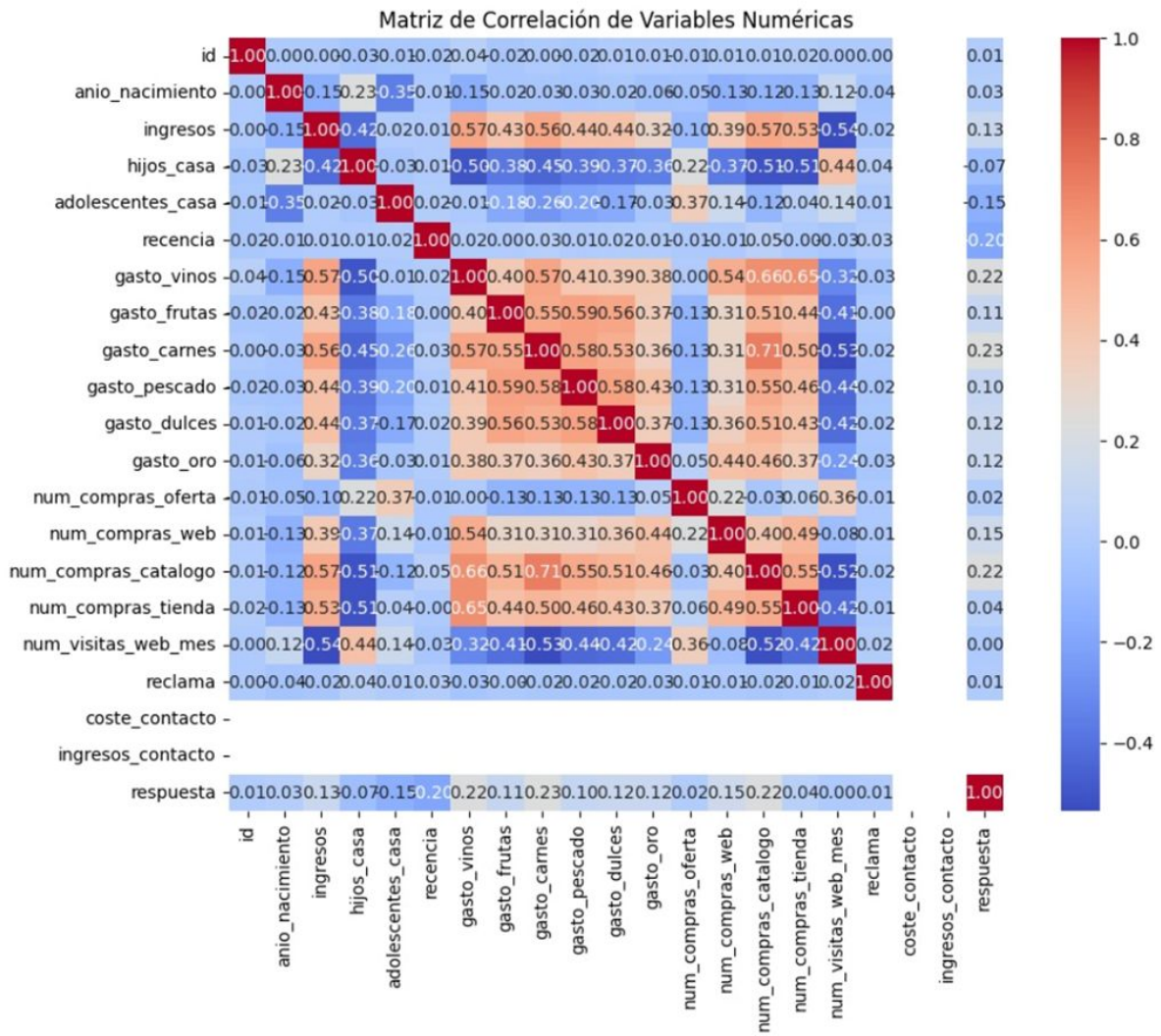
Variables Categóricas

Se detectaron variables categóricas que requerían estandarización:

- **estado_civil**: Valores inconsistentes
- **educacion**: Diferentes formatos

Distribución de Valores

Análisis Exploratorio de Datos



Preprocesamiento de Datos

Proceso de limpieza y preparación de los datos

Datos Raw — Preprocesamiento — Datos Procesados

Variables Sensibles

Se eliminaron todas las variables sensibles para proteger la privacidad.

- Nombres y apellidos
- DNI
- Números de teléfono
- Direcciones completas
- Números de tarjetas de crédito

Adherencia a buenas prácticas de ingeniería de datos y principios éticos de protección de datos.

Valores Faltantes

Estrategia conservadora de manejo de valores nulos.

- Se optó por eliminar registros nulos
- La pérdida de datos fue inferior al 10%
- Se mantuvo un dataset completo y consistente
- Se evitó la introducción de sesgos artificiales

Decisión basada en la mínima pérdida de información.

Variables Categóricas

Procesamiento para estandarización de variables categóricas.

- Identificación de variables categóricas
- estado_civil
- educacion
- Estándar de codificación
- Transformación para modelos ML

Preparación para modelos de aprendizaje automático.

Ingeniería de Características

Creación de nuevas variables para mejorar la capacidad predictiva de los modelos

Indicadores de Gasto

Ratios que normalizan el comportamiento de compra entre clientes.

- **compras_per_capita**: Compras por persona
- **intensidad_compra**: Frecuencia de compras
- **gasto_promedio**: Monto promedio por compra

Hábitos de Consumo

Variables que modelan las preferencias de consumo por tipo de producto.

- **pct_hist_carnes**: Historial de carnes
- **pct_hist_pescado**: Historial de pescado
- **pct_hist_frutas**: Historial de frutas

Interacciones con Campañas

Nuevas variables que capturan la historia de interacciones con promociones.

- **campanas_totales**: Aprobaciones anteriores
- **respuesta_relativa**: Proporción de respuestas
- **tiempo_desde_ultima**: Tiempo desde última interacción

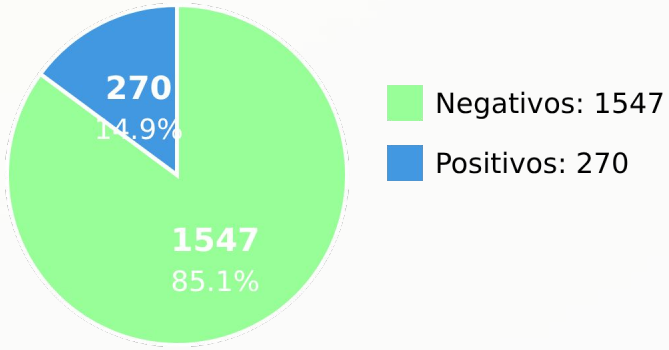
Ratios Demográficos

Mezclas como **ingresos_x_edad** que ayudan a comprender la capacidad de gasto real en función del momento de vida del comprador.

Problema de Clasificación

Predicción de propensión a aceptar campañas de marketing

Desbalance de Clases



⚠ **Desafío:** Solo el 14.86% de los clientes respondieron positivamente a la campaña (270 positivos frente a 1547 negativos).

Métrica de Evaluación

Recall
Sensibilidad

Precisión
Especificidad

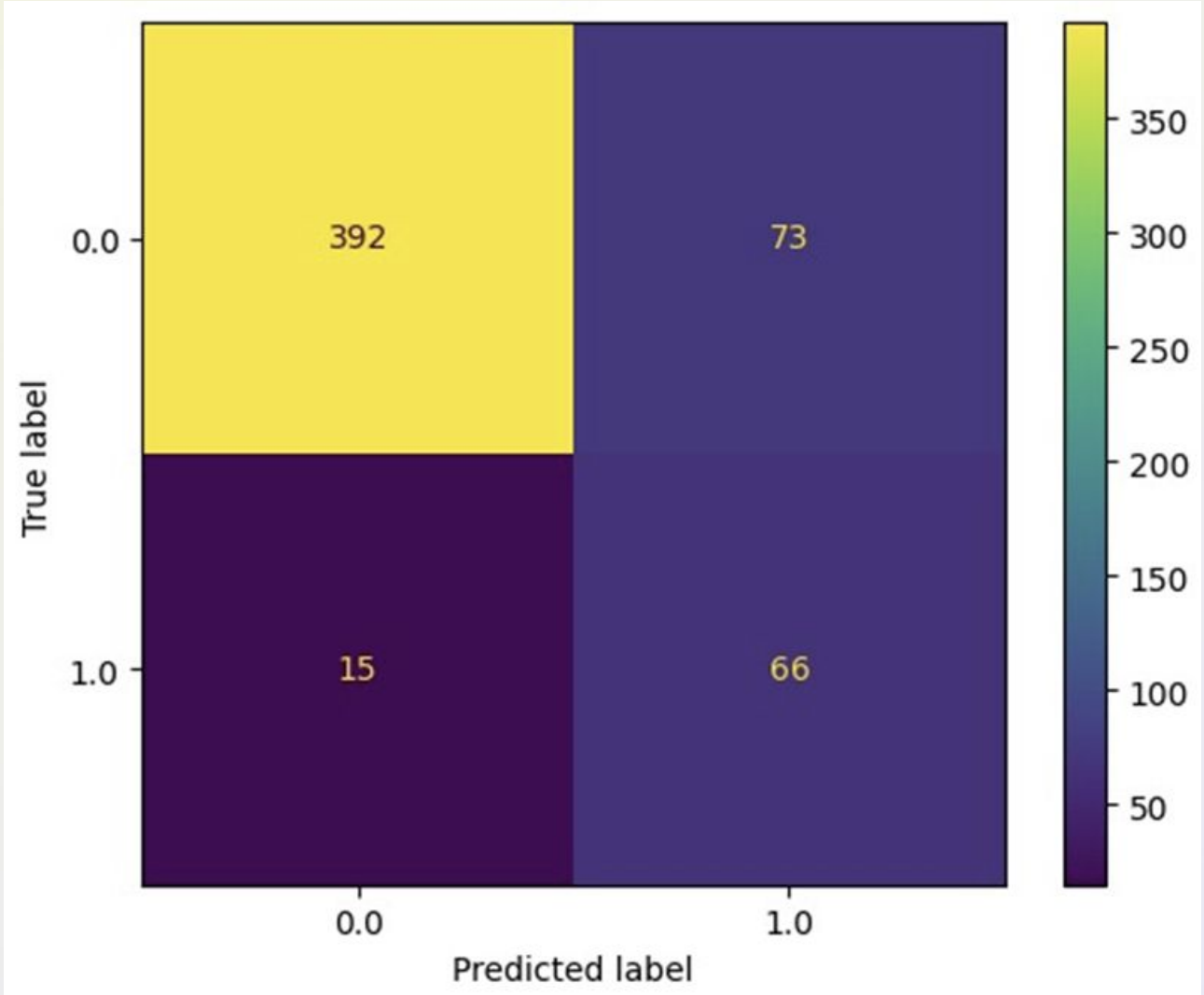
Justificación de la Elección

$ **Costo de falso positivo:** Bajo, limitado al coste de un email o SMS.

Costo de falso negativo: Alto, perder una oportunidad de venta.

Problema de Clasificación

Predicción de propensión a aceptar campañas de marketing



Estrategia de Clasificación

Priorización del Recall sobre Precisión

Justificación de la Estrategia

Equilibrio de Costes

- **Falsos Positivos:** Coste bajo (email/SMS)
- **Falsos Negativos:** Coste alto (pérdida de venta)

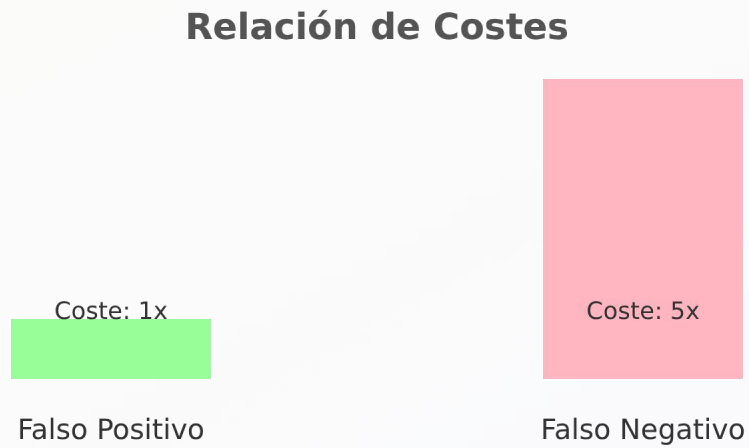
Enfoque del Modelo

- Identificar al mayor número posible de clientes interesados
- Aceptar cierto nivel de falsos positivos
- No perder oportunidades de venta

Resultado

Modelo optimizado para Recall, con una tasa de detección del 81.50% de clientes que comprarán.

Comparación de Costes

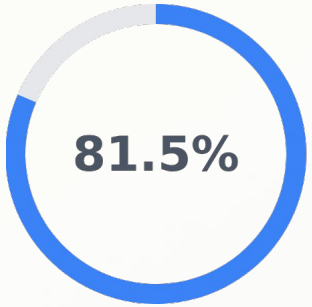


Priorizar Recall (sensibilidad) es la estrategia óptima dado el desequilibrio de costes.

Resultados de Clasificación

Regresión Logística con regularización L2 (c=10)

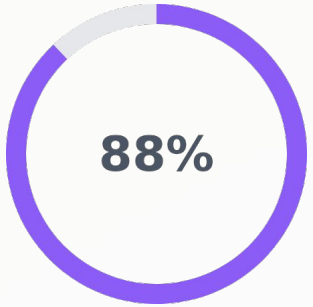
Recall



81.50%

Capacidad de detectar 8 de cada 10 clientes que comprarán

AUC-ROC



0.88

Buena capacidad para diferenciar entre clientes que compran y los que no

Accuracy

83.90%

Exactitud general del modelo a pesar de la focalización en la clase minoritaria

Precisión de la clase positiva: 47.5%. El modelo se diseñó para ser más agresivo en la identificación de posibles compradores, coherente con la necesidad de no perder oportunidades de venta.

Problema de Regresión

Objetivo

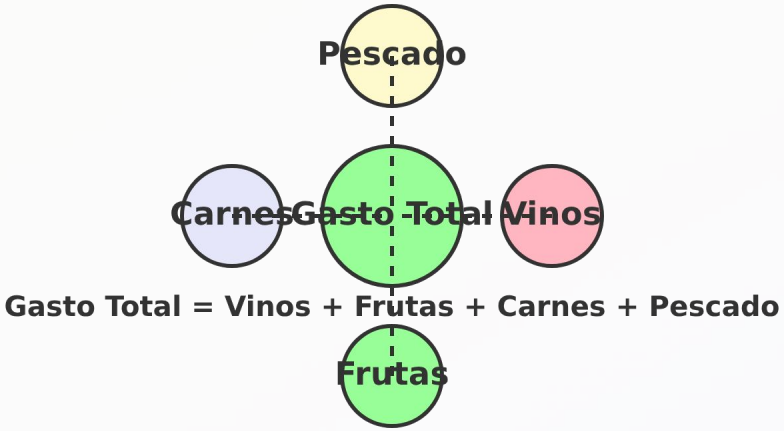
Calcular el gasto total anual estimado por cliente, para la planificación económica del supermercado.

⚠ Desafío: Data Leakage

- La variable objetivo era la suma directa de otras columnas
- El modelo arrojó un coeficiente R^2 cercano a 1.0

Esto haría que el modelo fuera inútil para predecir gasto de clientes nuevos

Visualización del Data Leakage



Impacto en el Modelo

- El modelo no podía generalizar a nuevos clientes
- Se necesitaba eliminar todas las columnas derivadas del gasto total

Solución de Data Leakage

⚠ ¿Qué es el Data Leakage?

Filtración de datos qué ocurre cuando la variable objetivo es la suma de otras columnas del dataset, haciendo que el modelo aprenda a sumar en lugar de identificar patrones predictivos genuinos.

> El modelo arroja un coeficiente de determinación (R^2) cercano a 1.0, indicando una predicción casi perfecta basada en la simple suma de variables.

Solución Implementada

- Eliminar todas las columnas que representaban el gasto directo
- Eliminar cualquier variable derivada de ellas (gasto promedio, porcentajes)
- Conservar solo columnas verdaderamente predictivas:
  - Datos demográficos
  - Actividad en la web

Selección de Variables

Variables Conservadas

- Demográficos
- Actividad Web
- Frecuencia Compra
- Otros predictores

Variables Eliminadas

- Gasto Vinos
- Gasto Frutas
- Gasto Carnes
- Gasto Pescado
- Gasto Huevos
- Gasto Promedio
- Porcentajes
- Suma Total

Resultados de Regresión

Modelo de predicción de gasto anual por cliente

Rendimiento del Modelo

El mejor modelo fue una Regresión por Máquinas de Vectores de Soporte (SVR).

Coeficiente de Determinación (R^2):

86%

El modelo captura una porción significativa de las fluctuaciones del gasto anual de los clientes, basándose únicamente en datos básicos de perfil y hábitos generales de compra.

Variables Predictivas Clave

Las variables con mayor peso predictivo del gasto anual incluyeron:

Factores ligados a las ganancias

Variables como `totalcompras_x_ingresos` y transformaciones cuadráticas (`ingresos_sq`), sugiriendo una relación no lineal entre la capacidad de compra y el gasto.

Maneras de comprar

Especialmente la frecuencia de compra utilizando el folleto en la tienda física, lo que mostró una fuerte conexión positiva con el gasto total.

Estos resultados demuestran que la antigüedad como cliente, la frecuencia de compra y el uso de medios clásicos (como el folleto o catálogo) son señales más claras de un alto gasto que la simple actividad en línea para este grupo de clientes.

Clusterización de Clientes

Segmentación de clientes para desarrollar tácticas de promoción diferenciadas

Algoritmo y Parámetros

Algoritmo seleccionado: K-Means

Número de clústeres: K=4

Criterios de selección:

- Separaciones claras y fáciles de interpretar
- Capacidad para identificar grupos naturales

Comparativa de Algoritmos

K-Means

- Clústeres bien definidos
- Resultados fáciles de interpretar
- Salidas estables y reproducibles

DBSCAN

- Forma grupos poco interpretables
- Clasifica muchos elementos como irrelevantes
- Inadecuado para divisiones variables

Visualización Conceptual de K-Means

Segmentos Identificados

Cuatro grupos de clientes con características distintivas

Padres Noveles

- Grupo más joven
- Ingresos por debajo del promedio
- Con niños pequeños en casa
- Hábitos de compra limitados a lo esencial
- Baja reactividad a las promociones

Buscadores de Ofertas

- Familias numerosas con ingresos ligeramente inferiores
- Muy conscientes del precio
- Expertos en el ahorro y uso de rebajas
- Ideal para campañas de grandes ofertas

Adinerados Digitales

- Individuos con carreras establecidas
- Buen nivel de ingresos
- Alta actividad en compras online
- Valoración de la conveniencia de la compra por internet

Compradores Valiosos

- Clientes maduros con los ingresos más altos
- Sin responsabilidades de hijos en casa
- Los que más gastan y mayor frecuencia de compra
- Muy receptivos a las campañas de marketing

Perfiles de Segmentos

Características detalladas de cada grupo identificado

Segmento 0: Consumidores Adinerados Digitales

Sección Web

- Individuos con carreras establecidas y buen nivel de ingresos
- Alta actividad en compras online, valoran la conveniencia
- Objetivo clave para programas de lealtad digitales

Segmento 1: Padres Noveles

Sección Esencial

- Más joven, ingresos por debajo del promedio, con niños pequeños
- Hábitos de compra limitados a lo esencial, baja reactividad a promociones
- Restricción económica que les impide aprovechar descuentos

Segmento 2: Hogares grandes buscando ofertas

Sección que mira el precio

Familias numerosas con ingresos ligeramente inferiores al promedio

- Muy conscientes del precio, expertos en el ahorro, utilizan las rebajas
- Ideal para campañas de grandes ofertas y promociones de volumen

Segmento 3: Compradores distinguidos más valiosos

Sección Exclusiva

- Clientes maduros con los ingresos más altos, sin responsabilidades
- Los que más gastan, los más fieles, mayor frecuencia de compra
- Muy receptivos a campañas de marketing, ideal para ofertas exclusivas

Implicaciones Estratégicas

Transformando el enfoque de marketing con estrategias personalizadas por segmento

Consumidores Adinerados Digitales

Grupo 0: Clientes con alto poder adquisitivo y actividad digital.

- Promociones específicas para internet
- Experiencia omnicanal integrada
- Programas de lealtad digitales

Fomentar la lealtad en línea con ofertas exclusivas digitales.

Hogares grandes buscando ofertas

Grupo 2: Familias numerosas con sensibilidad al precio.

- Grandes promociones y descuentos
- Reducciones de precios
- Campañas de volumen

Atraer con ofertas económicas y ahorros significativos.

Compradores distinguidos más valiosos

Grupo 3: Clientes maduros con los ingresos más altos.

- Mercancía selecta de alta gama
- Experiencias exclusivas
- Atención personalizada

Enfocarse en productos de lujo y experiencias premium.

Conclusiones

Cierre de presentación y conclusiones

Conclusiones Principales

Calidad y seguridad del dataset por la eliminación de datos sensibles

Nuevas variables relevantes que mejoran la capacidad predictiva

Se prioriza el recall para maximizar la detección de clientes potenciales

Se detectó y corrigió un caso crítico de data leakage evitando un modelo inválido

Se seleccionó K=4 por ofrecer una segmentación más accionable que K=3