Universidad Intercontinental de la Empresa

Memoria Proyecto Final Aprendizaje Automático

Nicolás Febrero, Sergio Arca, Alex Caride e Irene Valle

Contenido

1. Introducción y contexto 1
2. Análisis Exploratorio de Datos (EDA) 2
3. Preprocesamiento e Ingeniería de Características 3
   3.1. Gestión de la Privacidad y Limpieza 3
   3.2. Ingeniería de Características (Feature Engineering) 3
   3.3. Codificación de Variables 4
4. Línea de Trabajo: Clasificación (Propensión de Compra) 4
   4.1. Análisis Específico y Estrategia de Métricas 4
   4.2. Modelado y Gestión del Desbalanceo 5
   4.3. Resultados Obtenidos 5
5. Línea de Trabajo: Regresión (Predicción de Gasto Anual) 6
   5.1. El Desafío del "Data Leakage" (Filtración de Datos) 6
   5.2. Selección de Variables y Modelado 7
   5.3. Resultados Finales 7
6. Línea de Trabajo: Clusterización (Segmentación de Clientes) 7
   6.1. Selección de Algoritmos y Preprocesamiento 8
   6.2. Justificación de K=4 frente a K=3 8
   6.3. Caracterización de los Clústeres 8
   6.4. Caracterización de los Clústeres 9

1. Introducción y contexto

El objetivo principal de este proyecto es desarrollar soluciones de aprendizaje automático para una cadena de supermercados que está interesada en optimizar sus estrategias a la hora de la fidelización de sus clientes y sus estrategias de marketing.

Con base en el conjunto de datos proporcionado, hemos trabajado en tres líneas de actividad diferenciadas:

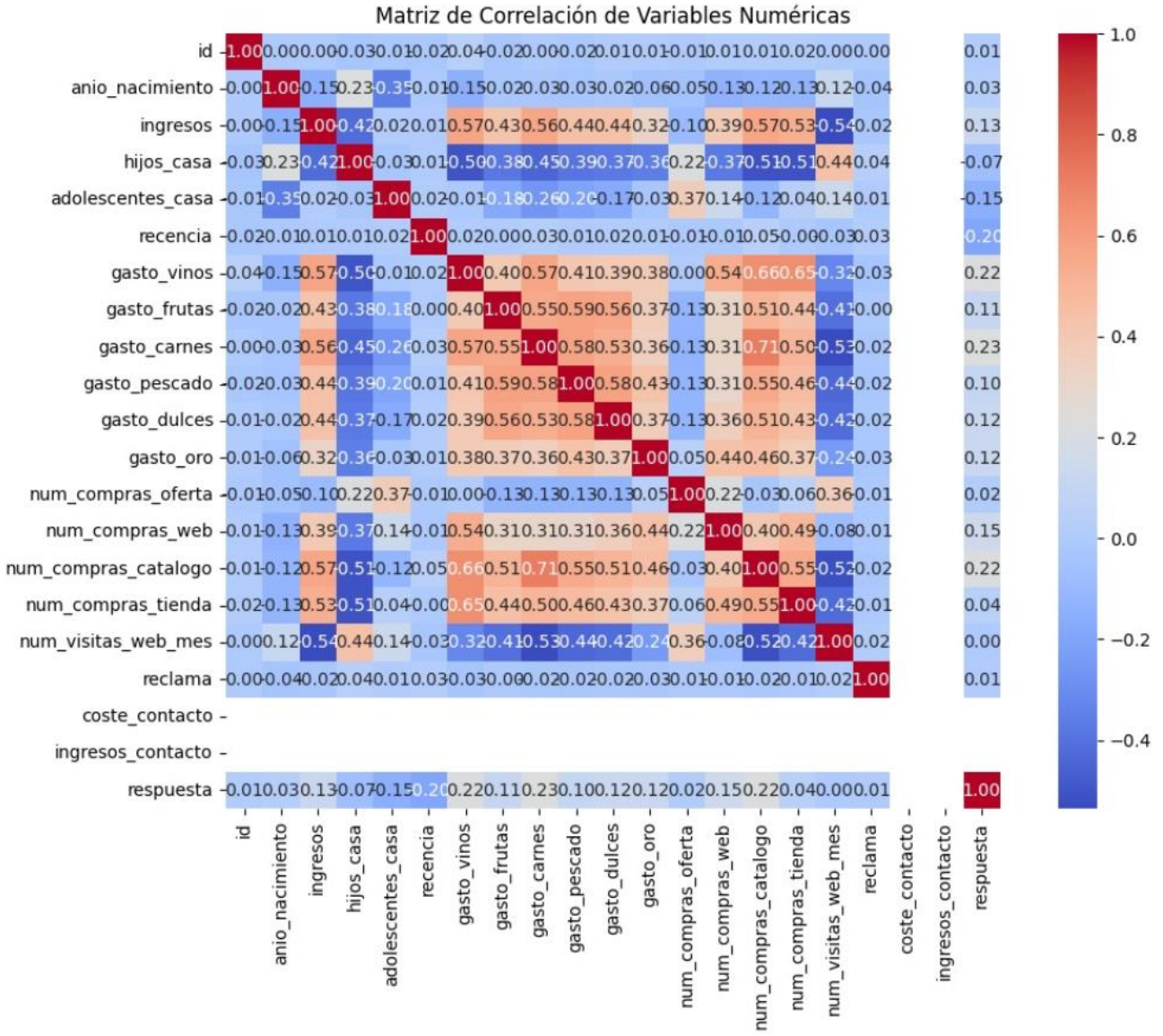
- La agrupación de clientes para definir perfiles
- La clasificación de la propensión a aceptar las campañas de marketing
- La regresión para predecir el gasto anual total de cada cliente.

2. Análisis Exploratorio de Datos (EDA)

Nuestra primera acción consistió en llevar a cabo un examen minucioso y a fondo del fichero que nos dio el maestro (proy_supermercado_dev. csv), con el fin de captar la estructura de la información existente y señalar fallos en su validez. Cuando importamos la información, notamos que había una colección de 1989 filas y 38 columnas, mezclando detalles de las personas, su rastro de gastos y aspectos de cómo respondieron a pasadas promociones.

Lo más importante que descubrimos fue que había datos personales sensibles: nombres, apellidos, documentos de identidad, números de teléfono, direcciones completas y también números de tarjetas de crédito.

Sumado a todo esto, al revisar los vacíos de información, notamos que cerca del 4 al 5 por ciento de las entradas tenían información ausente en campos cruciales como cuánto ganan, el año en que nacieron y su nivel de estudios. También identificamos variables categóricas como son estado_civil, educacion, que requerían estandarización para poder ser utilizadas correctamente en los modelos.



3. Preprocesamiento e Ingeniería de Características

Todo el flujo de preprocesamiento se implementó en el notebook preprocesamiento.ipynb, tomando decisiones para garantizar tanto la calidad de los datos como el cumplimiento de criterios éticos.

3.1. Gestión de la Privacidad y Limpieza

La primera decisión que tomamos fue eliminar todas las variables sensibles, es decir, nombre, apellidos, DNI, teléfonos, dirección y tarjeta de crédito. Esta acción sigue buenas prácticas de ingeniería de datos y respeta los principios éticos y legales de protección de datos, aunque sean simulados.

En cuanto a los valores faltantes, adoptamos una estrategia conservadora: eliminar los registros nulos en lugar de imputarlos. Dado que el perder datos era inferior al 10%, preferimos trabajar con un dataset

completo y consistente antes que introducir posibles sesgos mediante imputaciones artificiales en esta fase iniciales.

3.2. Ingeniería de Características (Feature Engineering)

Para conseguir incrementar el poder predictivo de los modelos, a parte de las variables originales, generamos nuevas características que representan patrones de negocio más complejos:

- ***Indicadores de gasto:*** las ratios como son compras_per_capita e intensidad_compra estas permiten normalizar el comportamiento de compra entre clientes.
- ***Hábitos de consumo:*** hay variables como pct_hist_carnes o pct_hist_pescado, ayudan a modelar preferencias de consumo por tipo y producto.
- ***Interacciones con campañas***: el dato total de campañas (campanas_totales) acumula las aprobaciones anteriores y sirve como un indicador claro para el objetivo de la clasificación.
- ***Ratios demográficas o proporciones de gente:*** son mezclas, como ingresos_x_edad, que ayudan a entender la capacidad de gasto real según el momento de vida del comprador.

3.3. Codificación de Variables

A las variables que no son numéricas las cambiamos para que los modelos de Aprendizaje Automático pudieran entenderlas:

- ***Codificación Ordinal***: en el dato de estudios, usamos una asignación numérica que respeta el nivel de formación.
- ***One-Hot Encoding***: en estado civil, generamos variables ficticias, por ejemplo estado_civil_casado o estado_civil_soltero, sin forzar una secuencia impuesta en un dato sin orden.

4. Línea de Trabajo: Clasificación (Propensión de Compra)

Nuestro segundo objetivo era predecir la variable objetivo respuesta: determinando si un cliente aceptará o no la oferta de la última campaña de marketing.

4.1. Análisis Específico y Estrategia de Métricas

En el notebook preprocesamiento_clasificacion.ipynb detectamos un desbalance significativo en las clases, tan solo el 14.86% de los clientes respondieron positivamente a la campaña (270 positivos frente a 1547 negativos). Ante este caso, tuvimos que decidir que métrica priorizar: ***Precisión o Recall.*** Para este caso de negocio, decidimos darle prioridad a la sensibilidad(recall) en vez de a la precisión.

El coste de un falso positivo, es decir, contactar con alguien que no comprará; es bajo ya que está limitado al coste de un email o SMS. En cambio el coste de un falso negativo, es decir, no contactar a alguien que si que comprará; es alto, esto implica que perdemos una oportunidad de venta.

Por ello, tomamos como objetivo identificar al mayor número posible de clientes interesados, aunque eso nos implique que tengamos que terminar por contratar a algunos que finalmente no terminarán comprando.

4.2. Modelado y Gestión del Desbalanceo

En nuestro notebook modelo_clasificacion.ipynb, probamos varios algoritmos, entre ellos los árboles de decisión y la regresión logística, usando la validación cruzada. Para abordar el desbalance de clases decidimos:

- No utilizar técnicas de sobremuestreo como SMOTE para evitar introducir ruido artificial.
- La configuración de los modelos con pesos de clase balanceados (class_weight = 'balanced').
- Aplicamos undersampling manual de la clase mayoritaria para reducir el desequilibrio.

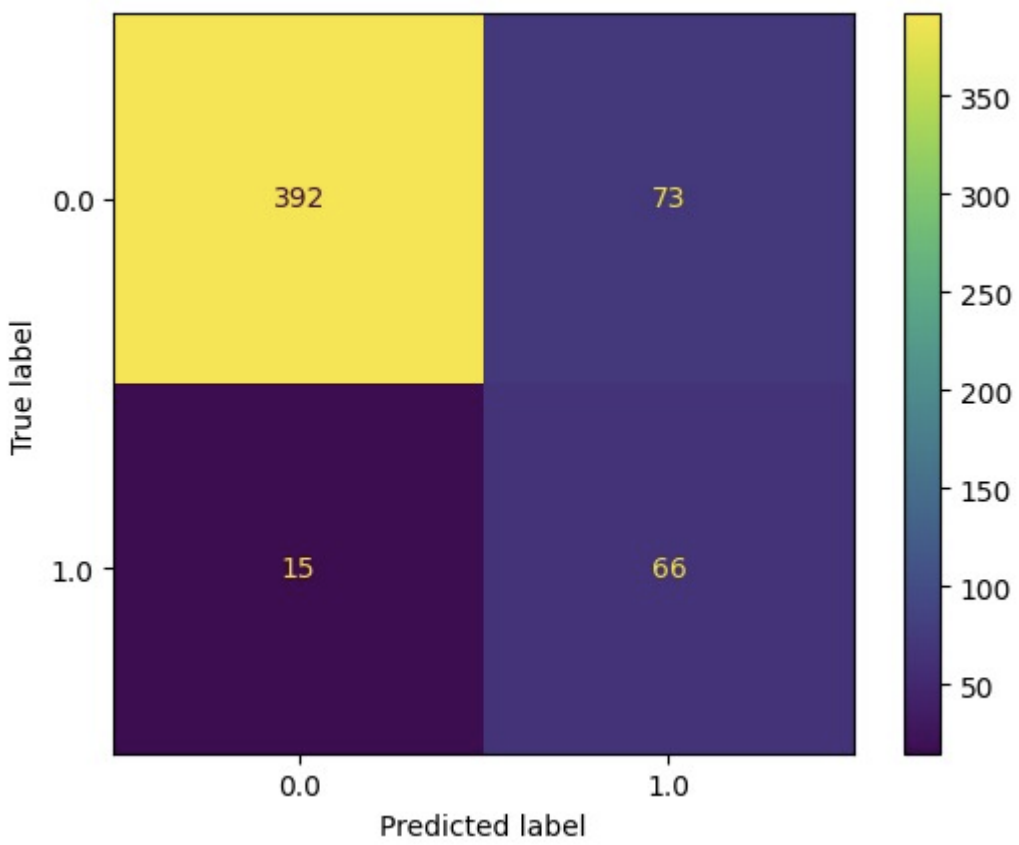
Así conseguimos que el algoritmo penaliza más los errores en la clase minoritaria, los clientes que compran, se terminan alineando con los negocios.

4.3. Resultados Obtenidos

Nuestro modelo final seleccionado fue una Regresión Logística con regularización L2 y parámetro c=10, integrada en un pipeline con escalado de datos. En nuestro conjunto de test, obtuvimos los siguientes resultados:

- ***Recall (clase=1): 81.50%*** → el modelo detecta aproximadamente que 8 de cada 10 clientes que compran y los que no.
- ***AUC-ROC : 0.88*** → indica una muy buena capacidad global para diferenciar entre clientes que compran y los que no lo hacen.
- ***Accuracy: 83.90%*** → conseguimos mantener una exactitud general elevada pese a que nos centramos en la clase minoritaria.

La precisión de la clase positiva es moderada (47.5%), pero termina por ser un efecto esperado de haber tomado la decisión de priorizar recall; el modelo termina por ser deliberadamente más agresivo identificando a los posibles compradores, algo que es coherente con la necesidad de la empresa de no perder sus oportunidades de venta.



5. Línea de Trabajo: Regresión (Predicción de Gasto Anual)

El tercer punto central del trabajo consistió en crear un sistema que pudiera calcular el Gasto total anual estimado para cada persona. Este dato es necesario tanto para la planificación económica como para la organización estratégica del supermercado.

5.1. El Desafío del "Data Leakage" (Filtración de Datos)

Durante las primeras pruebas, tuvimos un problema en común en el aprendizaje automático: traspaso de datos. Los resultados parecían ser excelente (un valor de R^2 casi de 1,0), lo que nos hizo dudar.

Cuando revisamos las relaciones entre variables, vimos que la variable que intentábamos predecir, el gasto total, era simplemente la suma de otras columnas en el mismo conjunto de datos (como puede ser el gasto en vinos, el gasto de frutas, en carnes, …).

Si dejábamos estas columnas al entrenar al realizar el entrenamiento del modelo, nuestro modelo solo aprendería a sumar, en vez de captar tendencias auténticas basadas en cómo es cada cliente. Con esto conseguiríamos que fuera inútil al aplicarlo a clientes nuevos de quienes todavía no tenemos el detalle de sus gastos por tipo.

Decisión Técnica: en nuestro archivo notebook llamado preprocesamiento_regresion.ipynb, quitamos todas las columnas que representaban el gasto directo y cualquier cosa derivada de ellas, como puede ser el gasto promedio o los porcentajes de gasto por grupo. Solo conservamos aquellas columnas que realmente sabíamos que podían predecir: los gastos demográficos, la actividad en la web y la frecuencia con la que compraban.

5.2. Selección de Variables y Modelado

Tras corregir el problema de la fuga de información, revisamos de nuevo las correlaciones para ver que factores predecían mejor el gasto. Las variables que más peso tuvieron fueron estas:

- ***Factores ligados a las ganancias***: tales como la totalcompras_x_ingresos o transformaciones cuadráticas (ingresos_sq), sugiriendo que la capacidad de compra y el gasto no tienen una línea recta.
- ***Maneras de comprar***: sobre todo cuántas veces compraba usando el folleto en la tienda física, mostrado una conexión positiva fuerte con el gasto total.

5.3. Resultados Finales

La mejor predicción la logramos con un modelo de Regresión por Máquinas de Vectores de Soporte(SVR) afinado mediante el ajuste de sus parámetros clave.

- ***Coeficiente de determinación en prueba (R^2) : 86%*** → el sistema captura una porción significativa de las fluctuaciones del gasto anual usando solo datos básicos de perfil y hábitos generales.

Respecto a entenderlo, el modelo demuestra que el tiempo que lleva comprando, la frecuencia y el uso de medios clásicos (el folleto o catálogo) son señales más claras de que alguien gasta mucho que la simple actividad en línea en este grupo de clientes.

6. Línea de Trabajo: Clusterización (Segmentación de Clientes)

Nuestra última línea de trabajo consistió en descubrir conjuntos naturales de consumidores mostrando conductas parecidas, buscando así la creación de tácticas de promoción distintas para cada uno de los grupos.

6.1. Selección de Algoritmos y Preprocesamiento

Puesto que esto era un asunto sin etiquetas previas, analizamos varias posibilidades. Aunque pensamos en usar métodos que se basan en la concentración, como puede ser DBSCAN, pronto lo dejaremos de lado.

Cuando lidiamos con mucha información de mercadotecnia, donde las divisiones entre grupos no son nítidas y las concentraciones cambian, DBSCAN suele hacer una de estas dos cosas:

- Juntar la mayor parte de la información en un único grupo grande
- Catalogar demasiados elementos como irrelevantes, desechando datos valiosos para el negocio.

Al final, decidimos usar K-Means ya que ofrece separaciones limpias y sencillas de entender, perfectas para clasificar clientes. Antes de eso, quitamos los valores extremos y normalizamos las variables para que el cálculo de las distintas fuera más fiable.

6.2. Justificación de K=4 frente a K=3

El determinar cuántos grupos debemos usar lo fundamentamos tanto en mediciones técnicas como en su aplicación práctica. La técnica del codo nos indica que tanto K = 3 como K = 4 eran elecciones sensatas.

No obstante, al usar K = 3 se formaba un solo "super grupo" de compradores con alto poder de compra que combinaba a gente con buenos ingresos con clientes muy acaudalados, tapando distinciones importantes en cómo actuaban, sobre todo en la vía digital.

Al movernos a K = 4, logramos una clasificación más detallada: el grupo de ingresos altos se partió en dos conjuntos diferentes:

- Clase alta establecida → grupo 3
- Usuarios digitales avanzados → grupo 0

Dividirlos es útil porque nos ayuda a diferenciar no solo en el momento en el que gastan, sino que también nos ayudan a diferenciar la forma en la lo hacen.

6.3. Caracterización de los Clústeres

Grupo 0: " Consumidores Adinerados Digitales" (Sección Web):

En este grupo están las personas que ya están establecidas en su carrera, con un buen nivel de dinero (0.71) y chicos adolescentes viviendo en casa. Las personas de este grupo son muy activas usando varios medios,

sobre todo internet (hacen 5.57 compras en línea). No usan tantas rebajas; valoran más lo fácil de comprar por internet que el costo. Funcionan bien para programas de lealtad enfocados en la tienda virtual.

Grupo 1: "Padres Noveles con Poco Dinero" (Sección Esencial):

En este grupo está la gente más joven, las personas que tienen sueldos más bajos que el promedio (-1.17 ajustado) y con niños pequeños (0.84). Las personas de este grupo compran poco y solo lo necesario. Casi no reaccionan a las promociones (0.09), lo que indica que no tienen margen económico para aprovechar descuentos.

Grupo 2: "Hogares grandes buscando ofertas" (Sección que mira el precio):

En este grupo están las familias grandes (con varios niños y jóvenes) y sus sueldos son un poco menores al promedio. Las personas en este grupo son expertos en ahorrar, usan las rebajas más que nadie (2.78) de todos. Necesitan estirar su presupuesto al máximo, por eso son perfectos para enviarles ofertas grandes y promociones como llevar dos por el precio de uno.

Grupo 3 "Compradores distinguidos más valiosos" (Sección Exclusiva):

En este grupo están los clientes que ya maduraron (1.09), aquellas con el mayor nivel de ingresos (1.31) y sin responsabilidades de hijos en casa. Las personas en este grupo son los que más gastan y los más fieles, estos tienen la mayor frecuencia de compra (0.95) y no les importa tanto el precio, pero son los que más responden a las campañas (0.49). Son los obvios para ofrecerles artículos de gama alta o exclusivos.

6.4. Caracterización de los Clústeres

Dividir el mercado en estas 4 agrupaciones significa dejar atrás la idea de un trato único para todos. En vez de mandarles el mismo tipo de promoción a todos los clientes importantes, la compañía ahora puede:

- Fomentar la lealtad en línea del Grupo 9 con promociones por internet y una vivencia integrada en todos los canales.
- Tocar la fibra sensible al costo del grupo 2 con grandes promociones y reducción de precios

- Enfocarse en el Grupo 3 ofreciéndoles mercancía selecta de alta gama.

De cualquiera de estos modelos el plan de mercadotecnia se vuelve mucho más verdadero y conectado con lo que cada clase de consumidor verdaderamente necesitan.